REPORT RESUMES

ED 003 401 24
IMPROVEMENT OF CLINICAL PREDICTION THROUGH SPECIAL TRAINING.
BY- KRAUSKOPF, C.J. AND OTHERS
MISSOURI UNIV., COLUMBIA
REPORT NUMBER CRP-S-120 PUB DATE 65
EDRS PRICE MF-$0.18 HC-$3.40 85P.

DESCRIPTORS- *PREDICTION, *MEASUREMENT TECHNIQUES, *COUNSELOR TRAINING, *FEEDBACK, MALES, COLLEGE STUDENTS, GRADE POINT AVERAGE, SCHOOL AND COLLEGE ACHIEVEMENT TEST, COLUMBIA, MISSOURI

A SERIES OF STUDIES WAS REPORTED ON INCREASING ACCURACY IN CLINICAL PREDICTION BY USING AN IMMEDIATE FEEDBACK PROCEDURE. THE GENERAL DESIGN IMPOSED EIGHT EXPERIMENTS AND THREE SETS OF ITEMS. SET ONE TESTED PREDICTIVE LEVEL BEFORE FEEDBACK. SET TWO WAS THE TRAINING SET. SET THREE TESTED THE EFFECT OF FEEDBACK TRAINING. THE PREDICTOR VARIABLE WAS TEST SCORES OBTAINED FROM THE SCHOOL AND COLLEGE ACHIEVEMENT TEST (SCAT "T"). THE PREDICTED VARIABLE WAS "PASS" OR "FAIL" BASED ON THE GRADE POINT AVERAGES (GPA) OF THE COLLEGE MALE SUBJECTS. THE PROBABILITY OF A CORRECT RESPONSE BY FEEDBACK TRAINING WAS INCREASED. IT WAS FOUND THAT COUNSELORS CAN MAKE MORE ACCURATE PREDICTIONS WHEN THERE IS IMMEDIATE FEEDBACK OF INFORMATION ABOUT THEIR ACCURACY. (RS)

U. S. DEPARTMENT OF HEALTH, EDUCATION AND WELFARE
Office of Education
This document has been reproduced exactly as received from the person or organization originating it. Points of view or opinions stated do not necessarily represent official Office of Education position or policy.

Improvement of Clinical Prediction Through Special Training

C. J. Krauskopf

University of Missouri, Columbia, Missouri

Project Number S-120

September 1964 to August 1965

ED 003401

IMPROVEMENT OF CLINICAL PREDICTION THROUGH SPECIAL TRAINING

C. J. Krauskopf
with
C. Imig
J. L. Williams
R. G. Neal
University of Missouri, Columbia, Missouri

Project Number S-120
September 1964 to August 1965

The research reported herein was supported by the Cooperative Research Program of the Office of Education, U. S. Department of Health, Education and Welfare.

TABLE OF CONTENTS

PAGE

PROBLEM ON WHICH THE RESEARCH WAS FOCUSED 1

Objectives 1

Related Research 2

Clinical versus Statistical Prediction Literature 4

Trends in the Literature 6

Contribution of Clinical Experience to Prediction. 13

Contribution of Training to Prediction 13

The Problem Investigated 22

PROCEDURES AND RESULTS 23

Preliminary Work 23

Exploratory Trial I 23

Exploratory Trial II 25

Experiment A 25

Experiment B 30

Experiment C 32

Experiment D 34

Experiment E 44

Experiment F 49

Experiment G 53

Experiment H-G 58

CONCLUSIONS AND IMPLICATIONS 68

REFERENCES 71

APPENDIX 77

LIST OF TABLES

TABLE PAGE

1. Intercorrelations among G. P. A., SCAT "T" and H. S. R. of Sets I and II 27
2. Mean Number of Correct Predictions Made by Six Ss on Task A Compared with Optimal Cutoff Lines and Multiple Regression Line 29
3. Results of Experiment B 31
4. Results of Experiment C 33
5. Means and Ranges of Distributions used in Experiment B, C, and D 39
6. Results of the Non-training Sets in the Major Runs of Experiment D 42
7. Results of Experiment E 46
8. Results of Experiment E_2 47
9. A Comparison Between the Distribution of Subject Predictions and the Actual Distributions of Experiment E_2 48
10. Comparison Between Means of N and T Groups on Each of the Four Scores 49
11. Results of Experiment F 51
12. Test Probability Matches in Sets I and III of Experiment F. . 52
13. Results of Experiment G_1 55
14. Results of Experiment G_2 56
15. Results of Experiment G_3 57

TABLE PAGE

16. Distribution of Approaches Used by the Ss in Experiment G_3 . . 58

17. Results of Experiment H-G 60

18. Distribution of Approaches Used by Ss in Experiment H-G as Determined From Written Statements 61

19. Prediction Study Summary Chart 62

LIST OF FIGURES

FIGURE PAGE

1. Scattergram of Items in Test with Three Predicting Diagrams 51

Problem on which the Research was Focused

The main question of this series of studies is: Can immediate feedback be used to increase accuracy in clinical prediction?

Objectives

If people can learn to improve the accuracy of their predictions about other people through training, the method would be of great use to counselors and educational administrators. Teachers, counselors and admissions officers continually make predictions about students, which are used by students and institutions with no knowledge of the accuracy of the predictions. It would also be of value if we could find out more about how people go about making predictions so that these could be communicated to others learning how.

As a secondary goal new information could be contributed to the statistical versus clinical prediction controversy if both clinicians (or naive subjects) and a regression equation could be trained and tested on similar data.

More specifically the following questions were asked:

1. Can counselors be trained to make more accurate predictions by training them in an immediate feedback situation where they learn the accuracy of each of their predictions immediately after each prediction is made?

2. Is there a difference in performance on this task due to clinical experience?

3. Is learning in this situation rote or is it governed by systematic rules?

4. What is the differential performance of clinicians trained by this method and a regression equation computed, or trained, on similar material and tested on the same material?

5. How complex can the relationship between the predicted variables and the simple two variable predictor be and still be detected readily by the naive subject when aided by immediate feedback.

Related Research

Prediction is an integral part of our life and language. As one involves himself in staff meetings at the professional level and conversations at the social level, it may become apparent that there are many statements predicting future relationships. These predictions may range from interpersonal predictions such as how one will act in response to a new hat, how one will act in response to treatment, or how one will react to the world of success and failure, to phenomena which are defined to be more physical, such as the outcome of certain methods of construction, yields of a variety of fauna and flora, or other such relationships. Efforts to speak without this prediction place a great inhibition on our ability to converse. One formal area of the study of this problem is prediction of behavior in psychology. This problem is receiving increasingly more space in texts and may become a major aspect of psychology. Gough's (1963) excellent review of clinical versus statistical prediction cites prediction as the problem in applied psychology.

During the years between 1947 and 1950, P. E. Meehl gave lectures and accumulated data on the accuracy of clinical and statistical (which he also calls actuarial) methods of prediction. The result of this was his book entitled *Clinical vs. Statistical Prediction*. It was his contention that the problem is to predict "how a person is going to behave" and the main question related to this problem was "in what manner should we go about this prediction?" (Meehl, 1954, p. 3). This definition of the problem (involving the word "versus") stimulated numerous articles in this area; however, this definition of the problem has focused the attention away from the question Meehl has raised about how a person is going to behave, to "with what we know now, what is the *best manner* to predict?" Out of this a race started. Stephenson (1962, p. 101) says, "the controversy, once started, goes on and no doubt will continue for a long time." It is easy to view the literature as a series of contests between clinicians and actuaries using progressively more complicated methods, each attempting to prove the merit of his favorite technique. So far, the generalizations from this item by item test of statistical and clinical techniques are very limited.

It is Gough's definition of the difference between clinical and statistical prediction that is used for the review of literature for this study. He says, "the defining distinction between clinical and actuarial methods is . . . to be found in the way in which the data, once specified, are combined for use in making the prediction. If the procedures, however complex mathematically, are in principle such that a clerk, or a machine, or anyone else could carry out the necessary operations and that the result would be the same in all instances, then the method is actuarial or statistical in the sense here being discussed. If the combination is done intuitively, if hypotheses and constructs are generated during the course of the analysis, and if the process is mediated by an individual's judgment and reflection, then the method is clinical" (Gough, 1963, p. 530).

Clinical versus Statistical Prediction Literature

In his book, Meehl reviews twenty empirical studies, none of which was adequately designed to test the clinical versus statistical prediction controversy. He concluded, however, (1954, p. 119) that "in about half of the studies, the two methods are equal; in the other half, the clinician is definitely inferior." In a later article, Meehl (1957) increased the number of studies reviewed to twenty-seven. The results here were even more in favor of statistical prediction. Oskamp (1961) reviewed eleven additional studies and concluded that over half of these favored the statistician. He also concluded that this tally revealed little because of questionable procedures used in the studies.

There have been three studies not reported in these comprehensive reviews of literature. One by Sydiaha (1959) attempted to predict judgments of Canadian Army Officers who actually recommended certain selected recruits for advancement. It was found in this study that the clinical prediction (although this was questionably clinical by Gough's definition because of item analysis performed on the clinicians' Q sorts) was better than the statistical prediction. A study by Grebstein (1963) indicates that the statistical and clinical predictions came out even. A third study by Cline and Richards (1962) concluded that the statistical methods were better when the task was to predict group norms, but clinical predictions were superior when predicting the individuals position within the group. Although these studies will not be entirely re-reviewed here, the overall batting average is twenty-four in favor of the statisticians, four in favor of the clinicians, and fourteen with equal results.

The predicting tasks of the studies favoring clinicians are:

1. Successful continuation in undergraduate study from questionaires and tests of entering freshmen (Apostolakas, 1957).
2. Success in psychiatric training from credentials, tests and interviews (Holt, 1958)
3. Future judgments of personnel offices by Q-sort (Sydiaha, 1959).
4. Intelligence from Rorschach protocols (Grebstein, 1963).

The predicting tasks of the fourteen studies indicating statistical and clinical predictions are equal are:

1. Psychotherapy response from MMPI (Barron, 1953).
2. Cadet success from test data and interview (Bobbitt & Newman, 1944).
3. College grades from high school rank, references and student statements (Chauncey, 1954).
4. Grades in navy electrical school from test scores and personal data and interviews (Conrad & Satter, 1945).
5. Pilot success from interviewer judgment (Dunlap and Wantman, 1944).
6. Clinical psychological trainees success from objective scores and credentials (Kelly & Fiske, 1951).
7. Cadet success from interview (Lepley & Hadley, 1949).
8. Academic success from tests, questionaires, and interviews (Sarbin, 1942).
9. Academic success from test scores and from personal data (Melton, 1952).

10. Recidivism from a prison doctor's knowledge (Scheidt, 1936).

The above ten studies are reviewed by Meehl (1954).

11. Organic brain damage from Bender-Gestalt (Goldberg, 1959).
12. Outcome of home care placement from abstracts of case histories (Ullman & Berkman, 1959).
13. Q-sort descriptions from abstracts from case histories (Weinberg, 1957). These three additional studies are reviewed by Oskamp (1961).
14. Predicting I. Q. from Rorschach protocol (Grebstein, 1963).

The criteria covered in predictive studies include:

1. Success in training or schooling.
2. Intelligence.
3. Recidivism.
4. Recovery from major psychoses.
5. Results of psychotherapy.
6. Brain damage.
7. Religiosity.

Because of the small number of studies in some of these areas, it is not possible to tell whether there is a significant difference in the content area and predictability (by clinical or statistical means). If one were to pick between these two on the proportion of studies showing its superiority, one would have to pick the statistical method.

Trends in the Literature

There seem to be two trends in the literature. One of them is exemplified by the title of the article by Holtzman (1960), "Can the Computer Supplant the Clinician?" He sets forth three controversies:

"the clinician versus the machine as a collector of data," "the clinician versus the machine as an interpreter of data," "the clinician versus the machine as a processor of data." He concludes, tentatively that the primary role of the machine could be as a processor of data. At this point he feels that the clinician has an edge over the machine in collecting and in interpreting information because of the broad range of possibilities and techniques available to the clinician. Stephenson (1962), Edwards (1961), Meehl (1957), Muldoon & Oakley (1958) and many others essentially support this general trend by suggesting computers take over areas of prediction where they can demonstrate superior function.

In 1959, Meehl's article offered some rumination on the validations of the clinical procedures. In it he reaffirms his position dealing with classifications of behavior. He says he believes that Kraepelin was not mistaken in his ideas of constructual categories, and that as a clinician, it is important to develop diagnostic skills in identifying or distinguishing classes of behavior (or rather nosological categories?). Meehl states, regarding a problem of prognosis, that there is as much evidence that one can predict a person's illness from diagnostic categories as there is that one can predict the course of illness from tests. The constructive proposals he offers include the challenge to the clinician to become a master diagnostician as well as to re-emphasize the statistical and qualitative treatment of all data. He believes that one should seek mathematical models to take over the routine activities pursued by psychologists and, in view of the immediate problem, to rely on statistical information gleaned from tests that are already given in the course of treatment. Meehl's conclusions could be summarized as follows:

1. There is a separate place or function for clinical and statistical prediction.

2. The statistician is the final authority on effectiveness.

3. Effort should be made to quantify and refine clinical data for statistical use.

4. The clinicians function by predicting membership in a behavioral class.

5. Methods for prediction must consider clinicians as a group.

6. Clinical intuition means making rather rational judgments based upon an accumulation of experience.

7. The logic of clinical activity necessitates having an objective accumulation of tabularized data.

The second direction is exemplified by Holt (1958). He attempts to re-define the research to show that clinical prediction in its own right can be as significant and as important as statistical prediction. He has criticized Meehl for creating an artificial dichotomy. The issue, he says, should more probably be stated as "the role of clinical judgment in predicting human behavior." Holt proceeds to give a model of logical structure from which we may formulate good clinical judgment. His five step process is as follows:

1. Acquaint ourselves with what we are trying to predict.

2. Decide intervening variables to be considered.

3. Determine data needed to measure intervening variables.

4. Gather and process data.

5. Combine process data to yield definite predictions.

The result of this process leads to a Type 3, Sophisticated Clinical Design. (Type 1 is called purely Actuarial, Type 2 is called Naive Clinical, and it is the Naive Clinical Design that is most often used in the studies of statistical and clinical prediction.) Holt continues "one kind of comparative study might teach us something, even though it would be hard to do properly; simultaneous attempts to predict the same criteria from the same data by clinicians and statisticians who have gone through the same preliminary steps." For example, statisticians will study the original group to determine the critical scores for the multiple cutting point (or whatever), the clinician will study the configuration of the scores of individuals of known performance; then, the respective techniques may be better compared. Holt did a study using Naive and Sophisticated Designs to improve the selection of medical men for specialty training in psychiatry. The results came out in favor of Sophisticated Clinical prediction.

Holt concludes, in a practical vein, that it would have been a mistake to concentrate on actuarial predictions to obtain a high and stable level of validity. This may tempt training centers to adopt a single mold in which to cast future psychiatrists. This method runs the risk of "confronting the future with a standard set of resources, derived from the past" (1958, p. 11). He admonishes the clinicians to refine their predictive procedures by "learning as much as possible about statistical prediction and adapting it to their own ends."

Hoffman (1960, 1962) draws a parallel between the mental process and the mathematical process. The problem he says is to determine and represent mathematically, the independent contribution of a series of predictive factors. This provides a way of giving the clinician a description of his mental processes. For example, the clinician describes the factors he is using and assigns them weight. Relative weights are then computed and compared with subjective weights and new weighting prepared that characterizes the individual clinician. In this manner, individuals can be compared with models most appropriate for making judgments. Further questions could be asked about the variability of the judge's personality and its relation to the ability to use such models. However, another approach is to teach the judge a more complex method of prediction. Specifically, can any judge be taught to make more efficient and more accurate use of information? Will this generalize or transfer to other judgmental situations (1960, p. 129)?

Hammond, et al, (1964) have adapted the Brunswik lens model to this end. Rather than having the judge describe his factors and weigh them, several "cues" are given him. These are correlated not only with the actual or "distal" fact, but with the judges' prediction. Thus, the amount of "weight" the cue provides the judge can be compared with its relationship to the criterion.

DeGroot (1961) entitled his article "Via Clinical to Statistical Prediction." After reading the article, it would seem that the more appropriate title would be "Via Statistical to Clinical Prediction." DeGroot intends to give the clinician the scores and rationale and the formula used for the numerical prediction and challenges the clinician to improve upon the formula's productivity. In addition to enabling the clinician to have a better start at prediction, the very fact that the clinician can beat the formula allows for a selective technique in hiring clinicians for certain predictive jobs. DeGroot admits that he should not like to recommend the use of this (he calls it improvement) design, as a test for clinicians because "of the weakness that it still has in common with the competitive designs, namely that it does not take into account the possibility of improvement of the clinician by experience and validation research on the job" (1961, p. 280). The real advantage he finds in this kind of improvement design is that it furnishes a realistic base line from which to measure the clinicians' contribution. Additionally, it can serve as a screening device for new hypotheses. As a model, it can also generalize to so many other disciplines. There are two additional advantages cited, however: one is that the design is a peace maker between the clinical versus statistical prediction controversy. It might be inferred that if the clinician attempts to use the statistical results and to improve upon them he may be less concerned with competing against the statistical design. The second advantage cited is that this is a model for hypothesis development. The rationale for this conclusion is: if a prediction is made from the actuarial formula "F", but improved upon by the clinician, then the clinician is accounting for a factor "Phi." The clinician in turn may try to account for "Phi" either by refining "F" (which is the statistical prediction), or by other variables. Apparently, DeGroot's method would allow the clinician to know when he should attempt to develop a hypothesis (because he has improved upon the statistical design), but does not furnish him any information on how to develop it.

Melton (1952) concluded that the "good" predictor, knowing the rationale, formula, and results of statistical prediction actually lowers his own predictive ability. Only if he is a very "poor" predictor does he improve his predictive ability. However, in no case did the counselor do as well as the statistical formula; consequently, this was not an adequate test of DeGroot's idea.

Thus it can be seen that this problem began with a concern about validity of method. The concern became vogue, partly because of the cliche which announced a contest between the statistician and the clinician. In some ways the results of this contest upset applecarts by "proving" that those who were claiming validity for their clinical predictions actually should take a back seat to methods which were very simple, far superior and much more reliable. For a time attention was focused more upon the outcome of the contest than upon the original problem.

Finally, the literature seems to indicate several directions which future studies could take.

1. The problem of prediction could be turned over to statistical methods and efforts to refine these methods be re-doubled.

2. Clinicians could incorporate statistical probabilities into their clinical judgments.

3. Clinical judgment could be modeled mathematically.

4. Clinicians could improve and methodize their clinical designs (After Holt, 1958).

There are two other possibilities that are slighted by the literature:

5. Identifying those clinicians who have superior ability to predict.

6. Training clinicians specifically to improve their ability to predict.

This study will concern itself with an exploration into these latter areas.

One conclusion that must be reached from the foregoing discussion is that the clinician does have some ability to predict. The question that should be raised is "How did the clinician develop this ability?" This learning of the clinician is treated lightly in the literature. The assumption is that being a clinician has with it the seeds of prediction. Most studies dealt with "experienced" clinicians - this meant they were not just students, but had had at least a year's experience.

Contribution of Clinical Experience to Prediction

The question then arises about the contribution of experience to prediction. This may be difficult to answer. Meehl feels that clinical intuition is a sum total of past experiences. Within the work experience, clinicians may be hesitant to make predictions. Indeed, it is often the administrative staff, faced with a decision to make, that consults the psychologist for principles to use in the matter. The bulk of the decision-making practice goes to someone else. When Melton (1952) studied counselors who were predicting first year honor points for entering freshmen (from high school rank and test scores), he found no steady rate of improvement, but rather an interaction between the variance of the counselors and weeks of experience. Counselors with previous experience did not do better than the others. Grebstein concluded that clinical experience beyond an introductory course in Rorschach did not significantly increase ability to predict I. Q. scores from Rorschach protocols. He indicated that this is not a function for which clinical psychologists are specifically trained, but one which they frequently do (1963, p. 132).

There is no information about the backgrounds of the clinicians in the studies reviewed. There is little evidence of any special training that would prepare them to predict. If the student clinician is to become predictive at all, it would be to predict what his instructor concludes would be the right answer. This training bears deeply upon what Meehl indicated was the correct approach - to diagnose or predict in a Kraepelinian fashion a class of behavior. The study by Holtzman and Sells (1954) may actually show the fallacy of this approach. Eminent clinical psychologists were asked to predict success in military flight training. None of the judges improved on chance. What is important, though, was that there was a significant amount of agreement among judges. This would seem to imply that the judges (independently trained) viewed a common classification as Meehl suggested, but the classification had no relationship to the criteria.

Professional journal articles do not seem to bear on this problem, and any concerted effort to improve in one's predicting ability almost has to be an individual effort, or the effort of a few colleagues pooled together. It could be concluded that this effort is seldom well organized or maximized. The conclusion is that there is no evidence of relationship between the amount of clinical experience and predictive performance.

Contribution of Training to Prediction

Insignificant differences between the trained and the untrained have been found by Hammond (1955), Goldberg (1959), Melton (1952), and Plaut & Cranwell (1955), using categories such as psychologist-non-psychologist, clinician-secretary, graduate-undergraduate, etc. Crow (1957) found that training medical students by instruction on patient-physician relationships after personal contact decreased accuracy in interpersonal perception. He concluded the little increase in validity was more than offset by an increase in the subject's variance (there was no "experience-trained" group). Studies by Cline (1955), Oskamp (1961), Newton (1964), and Thomas & Mayo (1957) tip the scale back toward the trained.

Cline (1955) found the general aspect of professional training enabled a better prediction of verbal and social behavior from filmed interviews. Since the three levels of professionalism were psychiatrists and psychologists, nursing trainees, and three undergraduate college students, there may have been other factors responsible for this professional difference. Newton, et al (1964) found the effective condition of training was knowledge of the relationship of the subject's variation, as well as the data relationships. Thomas and Mayo (1957) offered knowledge of correct and wrong predictions, presented a scatter plot of predictions and counselees' grades, compared individual counselor's accuracy with that of the group, and gave the counselor a chance to examine the records of the subject's wrongly predicted after the first group of subjects had received vocational placement. This long term feedback training was followed by substantial improvement in prediction on the second group of subjects.

Oskamp's (1961) study directly approaches the problem of training. In this study, three classes of judges - experienced clinicians, staff trainees, and inexperienced undergraduate students - were to judge whether MMPI profiles came from medical or psychiatric patients. The results show that the difference in performance between experienced and the inexperienced was significant beyond the .01 level. "Thus, it appears that this is a task which professional training and experience 'pay off' in significantly higher accuracy. However, the absolute magnitude of the difference between the average staff member and the average undergraduate is only about 3.5 percentage points, an amount which is relatively unimportant in terms of the number of correct predictions. In fact, the main reason for the obtained significance is the restricted range of accuracy scores among the clinicians, resulting in the small . . . standard error of the difference between means. The most surprising feature about these results is the excellent performance of the untrained undergraduate judges who had only a one and one-half page introduction to the MMPI upon which to base their predictions" (1961, pp. 75-76).

Oskamp then offered training for these inexperienced undergraduates and found that this training increased their level of performance up to the best actuarial methods. The training consisted of giving feedback after each set of profiles were completed. Some were given feedback in terms of accuracy, and some were given feedback in terms of appropriateness. The accuracy group, which we are mainly concerned with here, was given three bits of information which were called feedback:

1. Each trainee was given his score from the original set of profiles and from the just-completed profiles. His score was then compared to the scores from the rest of the class.

2. Each trainee was given a sheet with the four best actuarial systems described in full, plus the percentage that could have been expected had each system been used independently. This information was compared with their results.

3. Each individual in the training session was given correcting information after every ten test profiles and allowed to study their characteristics in an attempt to gain other helpful signs in predicting.

This, then, was the training which helped the inexperienced undergraduate to perform at the level of the actuarial prediction - between the level of "clinicians in general" and the "best clinicians." In order to gain some appreciation of the magnitude of this difference, some of these results are quoted:

1. Mean of clinicians.................................... 71.6%
2. Mean of undergraduates after training.................. 72.6%
3. Mean of four best signs (actuarial).................... 73.0%
4. Mean of best clinician................................. 73.6%

As an example, two of the actuarial methods taught the trainees were simple statements of probability such as, "four or more clinical scales 70 or higher, call p = .75," and "highest scale 76 or higher, call p = .72."

Thus, it is shown that training had an effect on the inexperienced judge, but apparently the training only taught them to do what the statistician does. This differs from telling the subject the actuary's criteria because he also received instruction in how to use it. Since the results equal the statistical prediction, the conclusion that the training session has "enabled the previously naive judges to perform at a level equal with that of the best actuarial method . . ." (1961, p. 96) should be interpreted strictly in light of this training.

The conclusions reached from the review of the clinical prediction literature indicate that some clinicians perform at a relatively high level on prediction tasks. On some tasks the clinicians as a group perform well. There is insufficient evidence as to the relationships between clinical experience training and areas of superior performance, but clinical training seems more related to predictive performance than does clinical experience. Additionally, specific training for the task by either providing knowledge of a model of results or study via "sophisticated design" has achieved greater results. In some cases brief training given to inexperienced subjects allows them to match the performance of the experienced subjects. What is needed now is an accumulation of evidence pointing to the relationship between experiences provided for the subject and the improvement of his ability to predict. If it can be shown that the subject can learn to improve his predictions, then the conclusions of the previous statistical versus clinical prediction studies are in doubt, but what is more important, new methods of clinical training and clinical experience may be indicated.

All of these studies (except Cline) attempted to teach the clinician to accept a model of data relationships. An additional question could be raised about the effect of feedback provided by instruction and definition of such a model. For instance, a model equivalent to the one in Oskamp's (1961) study, where "p = .75 if four or more scales are over 70," ought to bring a certain amount of "self-reward" on seeing and judging profiles, but it could be hypothesized that the more such models, the more confusing (and possibly less effective initially) the feedback system could become. Another kind of model is the paramorphic (mathematical) representation of one's mental process (Hoffman, 1960; Hammond, 1964). It is easy to see that this model for improvement may be very complex. Knowing which factors to give more weight may not be sufficient instruction to improve performance even though the instruction is valid. Some kind of additional training or practice may be needed in the assimilation of a new model.

There also may be many kinds of models leading to improvement in performance (some of which are better than others). Since it has been shown that the subjects of these experiments perform at a rather high level (some are colleagues) the models they may generate on their own initiative could be similar to the models found in the literature.

With these assumptions, there are two factors for consideration:

1. A better model than the existing constructs may be needed.
2. The dynamics of learning may be important no matter what scheme is used for clinical improvement.

Sarbin (1944) believes that the clinician attempts to develop some kind of an empirical frequency notion. This is clarified by saying, "Therefore, the clinician, if he is doing anything that is empirically meaningful, is doing a second-rate job of actuarial prediction" (Meehl, 1954, p. 31). If this is so, can the studies of the effect of feedback training on probability situations yield information to be used for the improvement of prediction? Generally these studies follow the paradigm suggested by Skinner: "The consequences of behavior may "feedback" into the organism. When they do so, they may change the probability that behavior which produced them will occur again" (1953, p. 59); "The organism must be stimulated by the consequences of his behavior if conditioning is to take place" (1953, p. 67).

Edwards (1961) studied the prediction of mutally exclusive events over a long series of trials. He concluded that people can learn to estimate probabilities in such a way as to maximize their correct prediction of outcomes. Stephenson and Zigler (1958) conducted similar studies with children and found responses of normal children were similar to the responses found in adults.

These and other studies of probability learning suggest that similar training techniques might improve prediction by clinicians in situations where probabilities are or can be known. It should especially be effective when the prediction asked for consists of two mutually exclusive propositions (Edwards, 1961). This model of probability learning could also meet the suggestions of Holt (1958), Oskamp (1961), and DeGroot (1961) in training for improvement in clinical prediction by:

1. Having clinicians and the formulas "trained and tested" on similar data.

2. Maximizing the effect of feedback by having it occur immediately after the response.

3. Providing a way for the clinician to express how he utilizes the data.

The major changes needed in adapting the learning to predict to a probability learning model can be enumerated:

1. Determine how much information is available to the clinician.

The problem should be chosen that allows the results of the data (feedback) to be as determinable as possible without being artificial. This would mean that for the investigation of the technique, data with available feedback knowledge is needed. There are several studies in the References dealing with validity of predicting college success from objective measures which are not reviewed here but are included in the references for sake of completeness.

2. Determine how to maximize performance of the clinician.

From the review of literature it is indicated that the subjects can learn to increase accuracy of prediction under conditions of delayed feedback of a model of relationships of the data. The question could be raised about the effect of simpler feedback. Could the subject's immediate knowledge of accuracy on individual items lead to an improvement in performance?

3. Determine how the clinician utilizes the data.

It should be possible to obtain descriptions of the methods used by clinicians and utilize these to discover methods used by the learning clinician.

There is a fourth aspect not previously discussed. This involves choosing an area where the clinician has previously demonstrated some ability. This should be an area possessing "face validity" so non-clinicians could be motivated. The area chosen for this particular study is success in college.

The Problem Investigated

Does accuracy in prediction improve as a result of certain experiences? One such experience postulated as holding promise is immediate feedback training. The general scheme of the study was to give the subject two numbers and ask him whether the student "passed" his first semester in college. These numbers represent an ability test score and high school rank, but this was unknown to the subject. During the training feedback was given by providing the actual status of the student after the subject responded, i.e., "yes" or "no". For the purposes of this study performance is defined as responding as to whether the student "passed," accuracy is the proportion of correct items; and learning is a positive change in accuracy between blocks of items.

Procedures and Results*

Preliminary Work

The initial exploratory investigations were used chiefly to iron out problems involved with task presentation and instruction.

Exploratory Trial I: Procedure: For the first pilot run, five subjects were used. These Ss were wives and friends of the experimenters. All but one had little or no previous idea of the object of the task. Thus, some idea could be obtained as to whether naive S's could understand what was wanted.

The task for this initial test required the Ss to predict whether a person passed or failed his first semester at the university on the basis of being told his score on the School & College Ability Test (SCAT "T" score). This score was based on the University of Missouri Norms.

The pass-fail criteria consisted of defining a first semester Grade Point Average of 1.8 or better as "pass". All scores below this were called "fail". The Ss' task was to see how many he could correctly categorize when given only the SCAT "T" score. The Ss were not told that the scores were SCAT "T" scores, only that they were scores obtained by each of the students they were asked to make a prediction on before the student entered university. The respective SCAT "T" and G.P.A. scores were randomly selected from those obtained by the entire freshmen male class registered for the fall 1959 semester at the University of Missouri.

A total of 135 scores were selected. These were divided into three sets containing 45 scores each. The first set served as a control to see how many scores would be correctly predicted before the feedback training. On the second set reinforcement was administered following the Ss' predictions. The S's were subsequently tested on the third set to see whether the training on the second set had enabled them to improve the accuracy of their predictions.

---

* Because this research consists of a series of separate but related studies procedure and results of each separate study are combined.

Each of the three sets had 24 fail and 21 pass items. Pearson Correlation coefficients between SCAT and G. P. A. scores for the three sets were 0.59, 0.59, and 0.63 respectively. The overall correlation coefficient was 0.70. The maximum number of correct predictions which could be made on each set were 33, 34, and 30 respectively. These maximums were based on a SCAT "T" score of 55 or less on Set I, 57 or less on Set II, and 56 or less on Set III being called "fail". If the cutoff score on Set III is made similar to that of Set II, which it should be if the Ss are making their prediction according to what was learned on Set II, the maximum number correct predictions possible becomes 29.

The necessary instructions were read to the S's before each set. Following the instructions, the E read the SCAT "T" predictor scores one by one, allowing the S ten seconds between scores in which to make a prediction. The S responded by marking "P" for pass or "F" for fail on a 3" x 2.5" card. Three sets of cards, each numbered one through 45, were given to each S. The reinforcements of Set II were on the backs of the cards provided for this set. The S made his prediction, then turned over the card to check his accuracy. Set II was administered immediately following Set I. There was approximately a 24 hour delay between the administration of Sets II and III.

Results: All five S's were able to understand the instructions and responded to the task satisfactorily.

Exploratory Trial II: Procedure: On the second pilot run, the major change involved the use of S's from the same pool that was to serve as a source for S's in subsequent work. These were five S's enrolled in Introductory Psychology during the fall 1964 semester at the University of Missouri. All such students were required to participate in three experiments as part of their course requirements.

Administration procedures were identical to the first pilot run except that the time allowed for the Ss to make their predictions was decreased. Five seconds was given in the first and third sets, eight seconds in the second.

Results: The mean number of correct predictions on Set I was 30.8. On Set II it was 30.2 and on Set III it was 28.6. At first glance these results seem to indicate that the Ss became poorer as the result of training. However, if the results are compared with the potential number correct when a consistent cutoff is used (33 on Set I and 29 on Set III), the Ss do show some improvement. They average 2.2 below the potential on Set I and 0.4 below that of Set III.

Prediction using a regression equation based on all 135 scores would yield 33 correct in the first set and 28 in the third. Thus the Ss were seemingly able to predict slightly better than did the regression equation though this difference would of course not be statistically reliable.

EXPERIMENT A

This task was in general similar to the preliminary studies. Once again Ss were asked to predict pass-fail from predictor scores. Among the major modifications made in the task were: (1) increasing the difficulty of the task through doubling the number of predictor variables from one to two, (2) reorganization of item make up on the tests to increase homogeniety across the three sets, and (3) running of the experiment twice instead of once. The last change was instituted as a partial test of limits.

Procedure: Two predictor variables were used for this task, High School Rank being employed in addition to SCAT "T". Scores on the latter variable ranged from 35 to 76, while scores on the former ranged from 37 to 80. G. P. A. pass-fail criteria was changed from 1.8 to 2.0. All scores 2.0 or greater were called "pass". These items were again randomly selected from those obtained by the freshmen male class registered for the fall 1959 semester at the University of Missouri.

Sets I and III of Experiment A were expanded to 50 items, and Set II to 100. Identical items were used in Sets I and III to make the results of the two Sets more comparable. Items of Set I were randomized to provide a different sequence in Set III. These same 50 items also formed half of Set II. The odd numbered items of Set II were those of Set I in a different sequence. The even numbers were new items. The pass-fail distribution was 48 percent pass, 52 percent fail on all three sets. Intercorrelations among the three variables of Experiment A, G. P. A., SCAT "T", and H. S. R., are given in Table 1.

Table 1

Intercorrelations among G. P. A., SCAT "T" and H. S. R.

of Sets I and II

		X_1	X_2	X_3
X_1 (G.P.A.)	Set I	-----	0.69	0.56
	Set II	-----	0.55	0.55
X_2 (SCAT)	Set I		-----	0.44
	Set II		-----	0.45

The multiple correlation coefficient of G. P. A. on SCAT "T" and H. S. R. was 0.75 for Set I and 0.64 for Set II.

The test was given twice, 24 hours between each running. Two different methods of presentation were used. The first involved use of cards and was very similar to the procedure used in the preliminary studies except that the test was administered by pre-recorded tape containing the instruction and the pairs of predictor variables for the three sets. Two pairs of sample scores were given before Set I and Set II to insure that the Ss understood the instructions. For the second run of the test, I. B. M. answer sheets were used (No. 1100 A 182 Rev.), the reinforcements for Set II being given via tape five seconds after each pair of scores. The Ss were asked to put a plus beside the item if they were correct. This was designed to help implant the reinforcements more securely. The make up of this tape formed the basis of all those made for succeeding experiments. Two voices were used. One gave the instructions and pairs of predictor variables while the other gave the sequence number of each item and the reinforcements of Set II. Ss marked their choices by the answer sheets by blackening under the "Y" for pass or under the "N" for fail.

At the conclusion of the experiment, the presiding examiner asked the Ss to write down on a piece of paper, both their feelings about the comparative merits of the two styles of presentation and a description of the way in which they went about making their decisions.

Results: Six out of the original eight Ss were present for both administrations of Experiment A. Table 2 represents the mean number of correct predictions for these six Ss on the two runs, together with the number of correct predictions using a regression line and the number using optimum cutoff points. A multiple regression equation was used to plot the regression line on a graph of the distribution of test scores. The optimum cutoff points were obtained by visual inspection of the graphed test scores. Set I cutoffs were; call "fail" when SCAT "T" was below 57 and H. S. R. was below 56, otherwise call "pass." For Set II cutoffs were; call below 56 on SCAT "T" and below 54 on H. S. R. "fail," otherwise call "pass." For Set III the potential number of corrects are based on the optimal cutoff lines and regression line of Set II.

No statistical tests were done on these results because it was apparent that no significant differences would be found. While the overall change is in the predicted direction there was no evidence that the second administration had any great effect. In fact, the Ss actually did slightly poorer after the repeat of the feedback training.

Table 2

Mean Number of Correct Predictions Made by Six Ss on Task A Compared with Optimal Cutoff Lines and Multiple Regression Line

	Set I (50)	Set II (100)	Set III (50)
Run I	35.33	69.16	36.66
Run II	37.33	69.32	36.00
Number of potential correct using optimal cutoff lines	42	80	40
Number of potential correct using regression lines	41	77	41

It was interesting to note that when the results of the 100 items of Set II were broken down into separate sets of 50, that the mean number of correct seen in Run I of the first 50 was 37.50 while that in the second 50 was 32.16. This was felt to be a result of the second 50 being more difficult then the first 50. That is, more of the pass-fail criteria did not correspond to the optimum cutoff points. As a check on this, the order on Set II was reversed on the tape made for the second administration. Once again the first given 50 came out higher than the second 50. The respective means were 37.66 and 31.66 correct predictions. It appears that the difference between the two 50's of Set II can not be attributed to differences in difficulty. The best alternative idea of why the difference occured seems to be that some Ss found fairly adequate cutoffs during the first 50 trials of Set II but were not satisfied with them because they still made errors and so experimented with other, less adequate cutoffs.

Judging from the written comments on task, only one S appeared to learn the optimum cutoff points, while two others guessed that they were around 50. One of the other three did not comment upon his selection procedure and the remaining two were quite vague about how they went about making their decisions. The Ss overwhelmingly preferred the second method of task administration. The vote was five to one in favor of the I. B. M. answer sheets.

EXPERIMENT B

Because of the small increase seen in the number of correct predictions of Set III over Set I in Experiment A, it was decided to increase the number of reinforcement items in Set II of Experiment B.

Procedure: The ten subjects obtained for this trial were from the Introductory Psychology pool. Task assembly and administrative procedures were identical to Experiment A except that Set II was increased from 100 to 150 items by the addition of 50 new items intermixed at every third position with the original 100 items. As in Experiment A, two runs were made.

Results: All ten Ss were present for both administrations of Experiment B. Table 3 contains the mean number of correct predictions by the Ss on both administrations. Set II is divided into consecutive groups of 50 for purposes of showing trends in the reinforcement set. Also included in Table 3 are the maximum number of correct predictions possible using the optimum cutoff score.

Table 3

Results of Experiment B

	Set I	Set II			Set II Total	Set III
		1-50	51-100	101-150	1-150	
Run I	35.1	36.9	37.2	34.3	108.4	38.0
Run II	36.2	38.0	37.9	35.0	110.9	37.3
Optimum cutoff	42				123	42

The optimum cutoff line is based on calling the item "fail" when SCAT "T" score is 56 or below and H. S. R. is 57 or below. Otherwise, the item should be called "pass." No regression predictions are given here. They will be given in the results of Experiment D which uses the same item data in a transformed state. The largest difference in test sets was between Set I and III of Run I. Run II apparently did not assist the Ss.

A t test was done on the above data comparing Set I and Set III of Run I. As an F test between the two variances involved was significant ($p < .05$), t was calculated as in case of unequal variances. The result was: $t = 2.2854$ which for 17 d.f. is significant at the $p < .025$ level in a one-tail test. An estimate of ω^2 was calculated to determine the amount of variation between Sets I and II which could be attributed to the feedback training (Hayes 1963, p. 327). The estimate of ω^2 was equal to 0.17. This suggests that a small to moderate amount of the variance was accounted for by the feedback training.

A review of the written comments obtained at the completion of the second running of the test indicated that the Ss benefitted little from the feedback training but rather relied more upon their previous experience with grades and averages to guide their predictions. All ten Ss described the experiment as boring and tiresome. This was mainly due to overlong pauses between items and having taken the test twice.

EXPERIMENT C

The results of the previous trials all suggested that something was needed to make the Ss rely more on the feedback training and less upon hypotheses gained from past experience. The tasks were evidently too much like what they were familiar with, thus they usually scored fairly high right from the beginning, the items of Set II contributing only a small amount to their decisions. For Experiment C the items of Experiment B were adjusted to try and force the Ss to get more out of the feedback training.

Procedure: Eight Ss were used for this trial. To alter the overall range of scores from that which had been used in the previous trials, all items of Experiment B were reduced by 20. The new respective ranges of SCAT "T" and H. S. R. became 12 to 56 and 17 to 60. There were few changes in administrative procedures. The major modifications were designed to increase motivation and interest in the task. On the second administration of the task, Ss were told their scores on Set I and III of the previous day and given the opportunity to score their answer sheets to see if they obtained better scores on the second day. Another change instituted was a reduction in the time allowed between items. Between item times in Sets I and III were reduced to three seconds. Set II times were lowered to five seconds; reinforcement was given three seconds after the item and there was a two second delay following this before the next pair of scores was given. The one other change in procedure involved switching the middle and last groups of 50 items in Set II from Experiment B to see if the slight observed drop in the last 50 in the results of Experiment B might be attributed to item difficulty.

Results: The results for the eight Ss who took Experiment C are given in Table 4. The optimum cutoff lines are 20 points lower than those of Experiment B. SCAT "T" becomes 36 or less and H. S. R. becomes 37 or less in order for an item to be called "fail".

Table 4

Results of Experiment C

	Set I	Set II			Set II Total	Set III
		1-50	51-100	101-150	1-150	
Run I	34.00	35.00	31.00	35.35	101.25	35.62
Run II	35.12	32.50	34.25	37.25	104.00	36.58
Optimum cutoff	42				123	42

A t test between Set I of Run I Set III of Run II yielded a t of 1.02 which is not significant at the 0.05 level. Calculation of $\underline{\omega}^2$ revealed that less than one percent of the variance was accounted for by the feedback training.

On the whole, the remarks written by the Ss at the conclusion of the second administration seem to indicate that they were less tired and bored than were the Ss who took Experiment B. As only two Ss took the opportunity to score their answer sheets on the second day in order to compare them with the results of the first day, it appears the reduction of boredom was chiefly due to the shortened between item time. The majority of the Ss used the reinforcement set to help guide their predictions. Evidently, adjusting the data achieved its purpose to some degree though the Ss did not show as large an improvement between worst and best sets of Experiment C as they did in Experiment B. Three of the Ss arrived at nearly optimum decision rules. Two gained little from the reinforced items. The remaining three benefitted only partially from the feedback training.

EXPERIMENT D

Further transformation was made in Experiment D to offset those who, during Set I, tended to take half the range as the pass-fail decision line. This time the actual distribution of scores altered to make the task more difficult.

After a preliminary run of Experiment D, a large scale test was undertaken. Administration problems had been largely ironed out by this time and it was felt that the feedback training technique was ready to be more thoroughly tested*.

Procedure: While most of the procedure of Experiment D has been recounted in the descriptions of the previous experiments, it will be given here in detail because it was done on a larger scale.

A trial run of Experiment D was given to seven Ss. The results looked promising and there were no administrational difficulties so it was decided to go ahead as planned. The Ss for the Imig study came from two sources. The first group came from an undergraduate psychology course. This sample was considered to represent a naive but intelligent population. Thirty-three students were present at both class sessions and successfully completed the two trials.

---

* The large scale test of Experiment D formed the basis of a Ph.D. dissertation undertaken by Charles Imig (1965). Much of the description which follows is taken from this dissertation.

The second source of subjects was from employed counselors in the Testing and Counseling Bureau of the University of Missouri, and advanced graduate students in the Counseling Psychology Practicum. This group was chosen because it was expected that the results could be generalized to similar groups in similar settings. This task was sanctioned as part of their official work load. These subjects had considerable training about college success and its relationship to High School Rank and SCAT scores, as well as having a general knowledge of the attrition rates of freshmen students. All had some work experience, and the general level of training was restricted enough to be found in almost any college counseling setting. There were 14 professional subjects who successfully completed the two trials.

One hundred and fifty items were randomly drawn utilizing the grade-point average, High School Rank and SCAT total score (freshmen male students, 1959-60). Out of these items, 50 were chosen at random and duplicated to become items for the pre- and post- tests.

The items were then recorded on audio tape and the subjects answered on I. B. M. answer sheets. In order to control for subtle voice cues from the tape, two people were used in the recording. One person gave only the scores of the SCAT and High School Rank, without knowledge of whether the items indicated pass or fail. The second person gave item number and "pass" or "fail" feedback during training.

The subjects received these instructions from the tape:

Pretest -- Set I

Narrator: "The purpose of this meeting is to see how well people can predict outcomes on the basis of two bits of information.

You will be given ratings which were obtained by college freshmen before they began college study. On the basis of these ratings we'd like you to decide whether the person obtaining them passed or failed his first semester at college.

Two ratings are given for each person. They are each two digit numbers.

Make your decision and mark it down with the pencil provided. In this set there are 50 responses. Blacken the "Y" for "Yes" if you think the student did pass and the "N" for "No" if you think the student did not pass.

You will be given three seconds to make your choice after you hear the two scores -- please blacken the "Y" for pass or the "N" for fail as the case may be.

Again the instructions in brief. You will be given a set of two ratings for each of 50 freshmen. On the basis of the ratings, make a decision as to whether each student passes or fails and blacken the appropriate space on the answer sheet.

Here's a sample to show you the routine. Use the sample card you have before you. Find Set I -- mark your decision -- "Y" for pass or "N" for fail in the space provided.

Ready, here's the sample: 30 - 51

During the pause, you should have blackened either the "Y" or "N". If you are unclear about anything indicate this to one of the proctors. Here's the first number."

Training -- Set II

Narrator: "There are 150 pairs of scores this time. There will be a short pause of a minute between numbers 50 and 51, and between numbers 100 and 101.

Here are two samples -- use the section on the sample card marked Set II.

Ready here's the first sample:

1. 60 - 60 (pause) Yes
2. 15 - 15 (pause) No

During the pause you would have made your decision and if your answer agrees with the answer given on tape you would have indicated this by putting a plus between the spaces for "Y" and "N" on the answer sheet. Are there any questions?"

Post Test III

Narrator: "This set is similar to the first set. You will be given 50 sets of scores. Blacken the "Y" or "N" space depending on your decision. You will not be told whether or not the person passed or failed.

Are there any questions?"

The subject was to mark the correct responses during the training period in order to maximize the awareness of the correct responses. The subjects were told they would receive the results of the first day's testing before the second test and if they were interested the final results would be discussed with them.

The tests were given twice to the undergraduates as a group during their regular Monday and Friday class period. After the second administration, they were asked to write a short paragraph about how they did what they did. The tests were given to the professional people individually in the testing room of the Testing and Counseling Bureau. The tape recorder, equipped with headphones was left in the room and the professional subjects were asked to schedule their time to take the test twice. The elapsed time between the first and the second test ranged from one day to two and one-half weeks. It was felt that the situation was not equally motivating for both groups, consequently an attempt was made to motivate the undergraduates class by a short warm-up session. The professional group received personal visits from at least two people, were asked at a staff meeting to participate, had the experiment considered as part of their work, and knew that a co-member of the staff could be aware of the results.

The previous studies indicated:

1. More than 100 items were needed in training.
2. The data, although chosen at random, was not normally distributed.
3. Time per item was a critical factor in motivation.
4. Small improvement occurred because subjects did quite well initially.

The method used in this study remedied these faults by:

1. Increasing the training session to 150 items.
2. Time per item was reduced to 6 seconds to maintain motivation.

3. Some of the additional items were picked to normalize the distribution.

4a. Twenty points were subtracted from each item.

4b. Items above 35 were increased by the difference between the item value and 35 (ex. 45 changed to 45 + (45-35) = 55).

The resultant distribution was denser in the lower third than in the upper two-thirds. Table 5, contains a comparison of the means and ranges among the item data of Experiments B, C, and D.

Table 5

Means and Ranges of Distributions used in Experiment B, C, and D

		SCAT "T"			H. S. R.		
		B	C	D	B	C	D
Mean	Sets I & III	52.90	32.90		51.88	31.88	
	Set II	53.19	33.19		53.26	32.26	
Range	Set I & III	35-76	15-56	15-77	37-80	17-60	17-85
	Set II	32-76	12-56	12-77	37-80	17-60	17-85

The subjects did well on the original pilot tests because they leaned heavily on linear relationships, and an efficient cutoff point for the two variables centered around 50. This was also close to one-half the range. This "naturalness" of the critical numbers provided enough cues to increase performance. The operations in 4a. and 4b. did not significantly change the relationships of the data, but did reduce the cues obtainable from the distribution (made the dichotomy based on one-half the range a less effective factor).

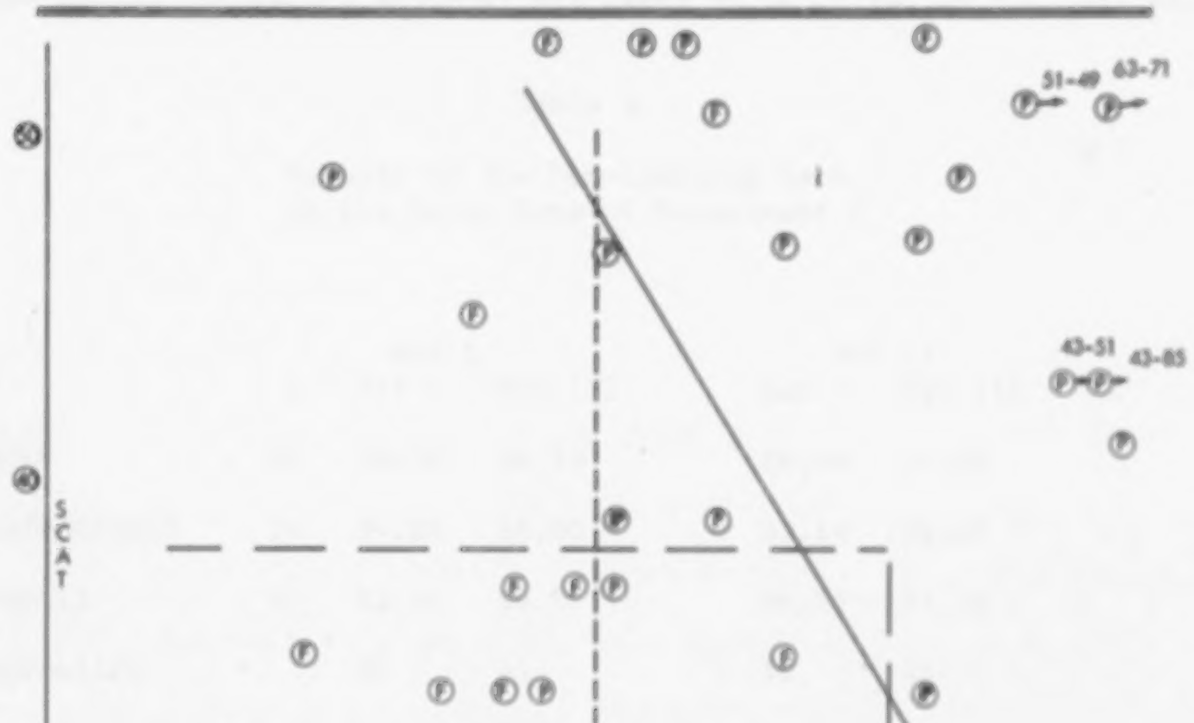
Results: In the trial run the seven Ss scored a mean of 30.57 correct on Set I and 35.86 correct on Set III. This difference was deemed large enough to indicate the Ss were having difficulty with the items of the first set and were probably using the reinforced set for information.

Perusal of the Ss' written statements showed that about five looked for cutoffs with most of these five getting quite close to the optimum. Of the remaining two, one guessed while the other attempted to remember specific pairs and their respective outcomes.

The results of the large scale run of Experiment D are contained in Table 6. The means are given individually as well as for both subject groups. A scatter plot of the data with cutoff lines and a multiple regression line plotted is given in Figure 1. The details of how the optimum number of correct items using (1) multiple cutoff and (2) a regression equation were arrived at are as follows: The multiple regression equation was based on a new sample of 150 items drawn from the same 1959 freshmen class and transformed in the manner described in the Procedure. The resultant prediction were not quite as good as a majority of the Ss were able to do. The analysis of the final scores of the subjects, indicates there are eight subjects in the undergraduate group and one in the professional group who scored 35 or less correct. This would mean that 81 percent of the subjects did better than the regression equation when they were "trained on similar data," and then tested on identical data.

The best predictions from consistent multiple cutoff lines could be made by (1) calling all scores which were both 37 or below on SCAT "T" and 39 or below on H. S. R., "fail." Using these cutoff lines would lead to

Figure 1
Scattergram of Items in Test with Three Predicting Diagrams



40 correct predictions. Overall, the Ss were not able to match this level of prediction, though they were not far from it. Four of the naive Ss and four of the counselors actually did reach 40 or better on the last set.

Table 6

Results of the Non-training Sets
in the Major Runs of Experiment D

		RUN I		RUN II	
	N	Set I	Set III	Set I	Set III
Naive	33	33.00	36.73	36.48	36.88
Professional	14	34.20	36.00	36.14	38.07
Overall	47	33.36	36.30	36.38	37.23
Regression		35	35	35	35
Cutoff		40	40	40	40

A *t* test on the results of the entire 47 Ss on Run II, Set III as compared to Run I, Set I gave *t* of 3.62 which is significant beyond the 0.005 level. The estimate of ω^2 based on this *t* is 0.11. Approximately eleven percent of the variance between these two sets is accounted for by the reinforcement sets.

To test the second hypothesis of this study, that the professional group would do better than the naive group, *t* tests were performed on both the first and last administered sets of 50. Neither *t* was significant at the 0.05 level. It does not appear that the professional group's experience helped them to any great degree on this task.

A similar test to the above was done on the differences between naive and professional Ss on the amount of improvement seen over the course of the experiment to see if the professionals profited more from the feedback training than did the naive group. The actual result was the reverse of this; the naive Ss' improvement was 4.22, while the professionals' was 3.79. However, this difference was not significant at the 0.05 level as determined by a t test.

The final hypothesis of the Imig study was concerned with how the naive Ss went about making their decisions. It was believed the Ss would not as a whole try to memorize pairs of numbers but would try and discover some principles to guide their predictions.

The test of this hypothesis was done by having the undergraduate subjects write how they formulated their answers to this experiment. These instructions were purposely kept unstructured in order to encourage the subject to put down his operating principle, rather than to ratify one suggested by the examiner. The statements of efforts to memorize specific pairs of numbers were sought in these explanations. Memorization was only mentioned by three subjects and this is mentioned in a minor manner (i. e. Subject 30: No real method. Usually a base of 75 sum total differential for all but about 15 combinations. These combinations had to be remembered). It can be surmised that these subjects attempted to remember false-positive items, or false-negative items. That is, they tried to remember the exception that did not fit their principle, which leads to rejection of this hypothesis.

These principles may be classified as "multiple cutoff score," "multiple regression," "combination of these," "differences," and "systems." Subjects are divided about equally among these first three while the difference model is used seldom and the systems approach is used apparently only once. Surprisingly the student who used this systems approach got a reasonably high score. This subject apparently detected something in the patterns of yes-no responses. The principle of using differences not not very pronounced, but some subjects mentioned that a large drop or difference in the scores had a particular or significant meaning. The multiple-regression principle was expressed by the subject's adding the scores together, and in some cases, giving a little more or less weight to the first score. Multiple cutoff scores were expressed by having subjects say that if both scores were above certain numbers, the subject passed, or if both scores were below a certain number, the subject failed. A combination of principles considered both scores above a certain number would be positive and both scores below a certain number would be negative, however, scores in either quadrant could be adjusted by an additive or multiple regression basis. (See Appendix).

EXPERIMENT E

This study marked a change from the direction followed previously which had culminated in Experiment D. The major modification concerned the variable to be predicted. The Ss were asked to predict whether a person who enrolled in engineering school stayed in this pattern or transferred to a different pattern. In addition, four predictor variables were used instead of two.

Procedure, Experiment E_1: Eight Ss were used from the introductory psychology subject pool. They were given one administration of this procedure.

The items selected for this experiment were taken from the records of students who registered in the first year engineering program in the fall of 1959. A total of 80 items were obtained. The four predictor variables were scores on the Missouri English Placement Test "T" Score (MEPT), (2) the School and College Ability Test Verbal, "T" Score (SCAT Verbal "T"), (3) the School and College Ability Test, Quantitative "T" Score (SCAT, Quantitative "T"), and (4) High School Rank "T" Score (H. S. R.). The 80 items were divided into two groups. The first and third sets contained the same 30 items in different orders. The remaining 50 items were used for the middle, reinforcement, set. Note that there was no item overlap between Set I-III and Set II. Approximately one-third of the items were "transfer" and the remaining two-thirds "non-transfer", in all three sets.

The instructions for each set of this experiment were given orally. I. B. M. answer sheets (Form I. T. S. 100 A 182 REV.) were used. The column headings on these "yes-no" answer sheets were blocked out and "N" and "T" were substituted. The Ss were instructed to mark under the "N" if they felt the person did not transfer out of engineering or under the "T" if they felt he did. The four predictor variables were administered by means of an 11 x 8 1/2" coil bound unlined notebook in which the groups of four scores had been inscribed in large numerals. Only one side of each page was used. The 11" dimension was the bottom of the page. Scores were arranged side by side and were clearly visible for a distance of 20 feet. The items of each set were put on consecutive pages starting with the last page and going forward. The E placed the book on a stand and flipped the pages over one at a time. The reinforcements of Set II were placed on alternate pages after their respective four predictor scores. The Ss were instructed that four seconds would be allowed for them to look at the scores and make their prediction. In the instructions for Set II, the Ss were told that following each item they would be shown whether or not they were right. If their predictions were correct, they were asked to place a "+" between the N and T blanks on the answer sheet. At the conclusion of Set II, written comments were asked for.

Procedure, Experiment E_2: This trial was identical to E, except that the seven Ss used, received the test twice, approximately 24 hours between each administration.

Results, Experiment E_1: Table 7 contains the mean number of correct predictions of the eight subjects and the percentage of corrects over total number of items for each set. The latter was included to enable comparison between Set II and the other sets.

Table 7

Results of Experiment E

	Set I	Set II	Set III
Mean number correct	15.50	27.5	16.25
Percent correct	51.67	55.00	54.17

There was an increase in Set III over Set I, however, this increase was slight (0.75) and all percentages are close to 50 percent. No statistical test of significance was done on this data. The written comments revealed a variety of approaches to the task, mostly involving the magnitude and/or the consistancy of the scores. The Ss did not appear to find useful, sharply defined methods of predicting. The experience provided by the feedback training did not influence the decision making process to any great degree.

Results, Experiment E_2: The mean number of correct predictions by the seven Ss who took both runs of E_2 are given in Table 8.

Table 8

Results of Experiment E_2

		Set I	Set II	Set III
Run I	Mean number correct	14.42	29.57	15.57
	Percent correct	48.06	59.14	51.90
Run II	Mean number correct	15.42	27.86	16.85
	Percent correct	51.40	55.72	56.17

The increase in correct number of predictions over the course of both administrations amounted to 2.43. About the same degree of increase was seen each day (1.15 and 1.43). At $\underline{t}$ test on the overall increment resulted in a $\underline{t}$ of 2.29 which is significant at the 0.05 level in a one-tailed test. The amount of variance accounted for by the two administrations of feedback training as determined by an estimate $\underline{w}^2$ was approximately 23 percent. Thus, while the Ss did not predict at a level much better than chance (50 percent), they did show a moderate amount of improvement apparently due to the effect of the feedback training.

The actual distributive of outcomes in this experiment was approximately 67 percent non-transfer and 33 percent transfer in each set. To see whether the Ss would probability match as a result of their experience on Set II, comparisons were made between the predicted and actual outcomes by means of Chi square tests. Table 9 lists the mean number of non-transfer predictions by the Ss on Sets I and III over the course of the two runs, as well as the χ^2 results of their comparison with the actual distribution.

Table 9

A Comparison Between the Distribution of Subject Predictions and the Actual Distributions of Experiment E_2

	Run I				Run II			
	Set I		Set III		Set I		Set III	
	N	T	N	T	N	T	N	T
Mean number of N and T prediction by subject	15.86	14.14	18.43	11.43	15.43	14.53	21.00	9.00
Actual distribution of N and T	20	10	20	10	20	10	20	10
χ^2	.683		.179*		.878		.000**	

* $p > 0.90$
** $p > 0.99$

These results show an interesting trend. The Ss divided their decisions nearly in half on both tries at Set I, but tended to probability match in each administration of the third set, especially on the last taken set where the probability is greater than 99 percent that the Ss were probability matching. The 24 hour interval between the two runs seems to have led to the Ss going back to a more even split in their decisions.

Comparison of the means of the N and T groups on each of the four scores, shows very little difference between them (Table 9).

Table 10

Comparison Between Means of N and T Groups on Each of the Four Scores

	MEPT	SCAT V	SCAT Q	HSR
Non-transfer	51.04	49.36	58.29	50.65
Transfer	51.14	51.64	57.38	50.00

This suggests that the improvement seen in the Ss' predicting ability was mainly due to better probability matching.

The written comments appeared to reflect a somewhat greater attention to clues offered by Set II, though a variety of methods was once again seen. The majority of the Ss payed attention to low and high scores, calling items with generally high scores, "non-transfer." One S tried to see how the last score related to the first one and called a relatively high last score, "non-transfer." Another subject predicted that highly divergent sets of scores would transfer. Two others commented that guessing and memory played a part in their decisions.

EXPERIMENT F

At this point we had to decide what to do with the limited time. We could proceed basically in two ways. We could complicate the predictors to see how people handled the increased information or we could change the predicted variable to make the relationship more complex to see how able the subjects were to detect more complex outcomes. Since preknowledge of the relationships between predictors and predicted variables is basic to this method we decided on the latter course. When the predictors are increased beyond two it becomes extremely difficult to know the exact extent of the relationship represented. Especially, since in our early experiments we found that linear regression did not adequately describe the usable relationship.

In this experiment we returned to a situation similar to Experiment D. Prediction was made more complex by asking the Ss to predict among five alternatives instead of between two as had been previously the case.

Procedure: From the standard subject pool, 35 Ss were obtained for this study and given one administration of the task.

The items for this test were the same as those used in Experiment D, thus the breakdown into sets was 50-150-50. The A, B, C, D, and E divisions were determined as follows: All G. P. A.'s above 3.0 were called A; those running from 2.3 to 2.9 were B's; C ran from 1.5 to 2.2; D from 1.0 to 1.4; and E included everything below 1.0. The resultant distribution of scores on Set II was seven percent A; 27 percent B; 37 percent C; 19 percent D; and 10 percent E.

Upon hearing each pair of scores, the Ss were asked to predict if the person got an A, B, C, D, or E grade in his first semester. These letters correspond to the column heading of the I. B. M. answer sheets which were used. (Form I. T. S. 100 A 444). As before, reinforcements were given in Set II and the Ss were asked, if they were correct, to mark "+" on their answer sheets immediately to the right of the space they have blackened.

Results: The data of Experiment F was analyzed in two ways: first, prediction improvement and second, distribution of prediction among the five choices. Table 11 shows the mean number of correct predictions on each set plus the mean number of predictions which were either correct or just one letter removed from the correct prediction, (e.g. if the correct answer was B, the S would get credit if he predicted A, B, or C. Also shown in Table 11 are the potential number correct using optimum cutoff and regression equations.

Table 11

Results of Experiment F

	Set I	Set II	Set III
Mean number correct	15.08	52.88	20.17
Mean number correct or one removed from correct	37.26	123.60	42.63
Potential correct from cutoffs	31		31
Potential correct or one removed	45		45
Number correct using regression equations	18		18

The Ss did improve significantly in their accuracy. $\underline{t}$ was equal to 5.10, which has a probability of less than 0.001 for 68 d.f. in a one-tailed test. An $\underline{\omega}^2$ calculation revealed approximately 26 percent of the variance was accounted for by the feedback training. A second $\underline{t}$ test, done on the mean number correct plus those just one removed from the correct answer, gave a $\underline{t}$ of 3.94 (53 d.f. because of unequal variances) which is significant at less than 0.001 in a one-tailed test. On this test, 17 percent of the variance was accounted for by the feedback training.

Table 11 shows that the Ss did considerably poorer than the optimum cutoff prediction. On the other hand, they did do somewhat better than the regression equation prediction.

The division into A, B, C, D, and E, and the associated varying probabilities of each category prompted a check into the Ss' probability matching in Experiment F. The correct probabilities based on the 150 items of Set II and the mean distribution of the Ss' predictions are given in Table 12.

Table 12

Test of Probability Matches in Sets I and III

of Experiment F

	Set I					Set III				
	A	B	C	D	E	A	B	C	D	E
Distribution of Predictions	1.77	5.83	15.03	17.17	9.60	1.97	11.08	21.08	12.83	2.97
Actual Distribution	3.67	13.33	12.33	9.33	5	3.67	13.33	12.33	9.33	5
X^2			17.6615				9.5517			
Probability			p <.01				p <.05			

Both Chi Squares show that the Ss predict distributions significantly different from the actual distribution. They apparently had much more difficulty in matching probabilities than did the Ss in Experiment E, though the Ss in the present study did improve considerably in their probability matching, as shown in the smaller χ^2 of Set III.

Written comments were obtained from only a few of the Ss who took this test. Of the eight who did comment, five took an average of the two scores then used cutoff points of these averages. This would appear to correspond to the "regression" method as outlined in the results of Experiment D. Two Ss treated the numbers as if they were consecutive and felt that if the second number was higher than the first, the grade would be high and vise versa. One S, having formulated no approach to the task, just guessed.

EXPERIMENT G

Prior to this trial, all our work had been based on linear relationships between the predictor and predicted variables; higher scores meant higher grades. However, correlation curves are not necessarily linear. Interesting non-monotonic relationships have been found between and among variables. This is illustrated by work with such areas as the relation of manifest anxiety to learning (O'Connell, 1965) and emotion to cue utilization (Easterbrook 1959). For example, extremes of low or high manifest anxiety seem to lead to more difficulty in learning than does a moderate amount.

Most previous studies on prediction have concentrated on linear relations though there have been a few which were non-linear. (Hammond, Hursch & Todd, 1964). Experiment G was designed to see if feedback training could lead to learning of other than linear relationships. For this purpose a test was designed which involved an inverted U distribution.

In other words, extreme scores at either end of a continuum had a high probability of the same outcome and this outcome differed from that related to the majority of the scores between the extremes.

In the course of Experiment G, three separate investigations were done, subtitled: G_1, G_2, and G_3. These differed only in the size of the reinforcement set which was increased on successive forms.

Procedure, Experiment G_1: Nine introductory psychology Ss were used and given one administration of G_1. The items for this test were taken from a random sample of records of males who were freshmen in the 1959 fall semester at the University of Missouri. SCAT "T" served as the predictor variable. It was to be used to predict either pass or fail. Sixty items were selected for use, 30 of which served to make up Set I and III (in different orders). All 60 were used in Set II, thus 30 of the items were present in all three sets. The pass-fail point was altered from the formerly used 1.8 to 2.0 range. Two points were selected. These were chosen so that they would give an evenly balanced inverted U distribution. All first semester G. P. A.'s below 1.30 and above 2.25 were made fail items, G. P. A.'s from 1.30 to 2.25 were pass items. This gave a predictive basis where extreme scores tended to lead to identical results which contrasted with scores in the middle of the distribution. The apportionment of scores which resulted from these cutoff points was 16 in the lower fail level, 30 pass, and 14 in the upper fail level.

The instructions and test items were presented via tape. I. B. M. answer sheets were used (Form I. T. S. 1100 A 182 REV.). Reinforcements were administered in Set II, and Ss asked to place a "+" between the Y and

and N blanks if they were correct.

Procedure, Experiment G_2: This procedure was identical to that of G_1, except that Set II was increased to 120 by repeating the 60 items of G_1 twice. Five Ss were given one administration of G_2.

Procedure, Experiment G_3: Both naive and experienced Ss were used. The former consisted of 35 Ss taken from the introductory psychology group. The latter were 28 college counselors who were taking a summer institute course at the University of Missouri.

Further expansion of Set II was done by adding 30 new items, randomly interspersed through the 60 items from Set II, G_1. These 90 items repeated twice, gave a total of 180 reinforcement items.

Results, Experiment G_1: The mean number of correct predictions on the three sets of G_1 are listed in Table 13. Also given are the potential number correct if optimum cutoffs are used. These optimum cutoff points were: call all scores from 51 through 57 "pass", and all others, "fail".

Table 13

Results of Experiment G_1

	Set I	Set II	Set III
Mean number correct	14.22	36.78	16.44
Potential using optimum cutoff points	25	48	25

No statistical test was made on the difference between Sets I and III. The Ss did improve somewhat, but were considerably below the potential obtainable if they all had discovered the inverted U shape and optimum cutoff points.

Written comments from the Ss revealed that the majority (six out of nine) relied upon linear predictive methods. Only one discovered the need for the double cutoff points, though he was not able to pinpoint them exactly. Two appeared to follow no specific principles in making their decisions.

Results, Experiment G_2: The mean number of correct predictions by the five Ss who took G_2 are in Table 14.

Table 14

Results of Experiment G_2

	Set I	Set II	Set III
Mean number correct	18.00	74.20	17.80
Potential correct using optimum cutoff points	25	96	25

The Ss actually did slightly worse on Set III than on Set I. Strangely enough, the written remarks of the Ss reveal that at least one was able to discern the exact cutoff points while two others at least realized two cutoffs were involved. One S tried to remember specific scores and their associated predictions. The remaining S guessed.

Results, Experiment G_3: The results of G_3, a major test of feedback training with non-linear distributions, are given in Table 15. The t tests contained in Table 15 are based on comparisons between Set I and Set III of the respective subject groups. Both groups show a significant improvement. No test was done to compare the two groups because of the small differences seen in the results of Set III. The estimates of ω^2 indicated that for the naive Ss, 34 percent of the variance was accounted for the reinforcement training while only 10 percent was accounted for in the case of the experienced group. It appears that the experienced group obtained less information from the feedback training than did the naive group. Perhaps stronger preconceptions about the relationship involved led the experienced Ss to do this.

Table 15

Results of Experiment G_3

		Set I	Set III	t
Mean number correct:	naive	15.34	18.91	6.03**
Mean number correct:	experienced	15.21	17.93	2.68*

*p <.01
**p <.001

Unfortunately, written statements about the test were not obtained from all of the naive Ss. Of the 28 out of 35 which were obtained, twelve actually realized that two cutoff points appeared to be involved. Three more were at least partially aware of this. Of the others, four used linear methods, three tried to memorize the answer for each score (retrieval), one guessed, three tried to be systematic (e.g., a sequence of 3 fail then one pass), and two were vague. Turning to the experienced group of 28 Ss, the division was as follows: seven discovered the inverted U to at least some degree, six used linear methods, one tried to retrieve, two employed sequences or systems, nine guessed and three were vague. The breakdown of the two groups is given in Table 16.

Table 16

Distribution of Approaches Used by the Ss in Experiment G_3

	Inverted U	Linear	Retrieval	Sequence	Guess	Vague
Naive	15 (54%)	4 (14%)	3 (11%)	3 (11%)	2 (7%)	1 (4%)
Experienced	7 (25%)	6 (21%)	1 (4%)	2 (7%)	3 (11%)	9 (32%)

The distribution in Table 15 supports the statements made above to the effect that the naive Ss were better able to profit from the feedback training than were the experienced Ss. A very peculiar finding was the great difference between the two groups under the 'vague' category. A considerably higher percentage of the experienced Ss seemed unable to put down on paper what system they used to base their predictions upon.

EXPERIMENT H-G

Many of the Ss who experienced the various trials of Experiment G were unable to discover the optimum cutoffs to base their decisions upon. Experiment H-G was designed to see if the adherence to previous experience could be loosened so as to enable them to gain more from the feedback training. Another test was given prior to the administration of G_1. This test was designed to prepare the Ss for an unfamiliar distribution by giving them experience with another distribution, the relations of which were thought to be easy to discover though quite different from the usual. As a control, another group of subjects was given the same inverted U distribution, but preceded by a straight forward linear distribution instead of the unusual one given to the experimental group.

Procedure: Two groups of subjects were obtained for this experiment. The experimental group had 32 Ss and the control had 41 Ss.

The experimental group of 32 Ss was first given a test where the relationship between predictor and predicted variables was linear but reversed from the usual order. High scores tended to fail, while low scores tended to pass. A single predictor variable, SCAT "T", was used. All associated first semester G. P. A.'s 2.00 or above were called "fail", those below, "pass". Sixty items were taken from the records of men students entering into the freshman classes, fall 1959, at the University of Missouri. Thirty of these items were taken for Set I - III. All 60 were included in Set II. Score distribution was: 31 pass and 29 fail in Set II and 14 pass and 16 fail in Set I - III. The control group was given scores identical to those received by the experimental group, however the usual straightforward relationship held; high scores tended to "pass" and low scores tended to "fail".

The construction of the new tests for the experimental and control groups (H_{rev} and H_{strt} respectively) was carried out in the conventional manner. Administration of instructions and items was by tape. Reinforcements were given in Set II. I. B. M. answer sheets were used (Form I. T. S. 1000 B191). Subjects were instructed to blacken the space under the T if they felt the person passed and under the F if they felt he failed. After the H sets had been administered, the Ss were given G_2 and told that the G sets would not necessarily involve the same methods of prediction which were suitable for H.

The form of the inverted U distribution used for experiment H-G was G_2 which had a 30-120-30 item distribution for the three sets. G_3 was not used because despite the fact that previous test showed it resulted in more learning than G_2, it was felt to be too long (240 items) to be given after the 120 items of the H series.

Results: Table 17 contains the mean number of correct predictions by the two subject groups as well as the results of *t* tests between Sets I and III within each test.

Table 17

Results of Experiment H-G

	H			G		
	Set I	Set III	*t*	Set I	Set III	*t*
H_{rev}-G_2	9.09	14.97	5.02***	14.82	16.03	1.60
H_{strt}-G_2	20.78	22.54	2.26*	15.80	17.12	2.62**

*$p < .025$
**$p < .01$
***$p < .001$

All tests in Table 17 are one-tailed. Comparison between G_2, Set III results of the two groups shows no significant difference (*t* = 1.45). These results were not as predicted. The subject group which experienced an urusual distribution prior to the G_2 did not make a higher mean number of correct predictions on Set III than did a group who experienced a linear distribution prior to G_2, the inverted U distribution. In fact, the H_{strt} group showed a greater improvement in the course of G_2 than did the H_{rev} group, although only seven percent of the variance in G_2 of H_{strt} was accounted for by the feedback training. As no significant difference was seen between Sets I and III of G_2 in the case of the H_{rev}. group (only 2 percent of the variance accounted for by the feedback training) the experience of H_{rev} does not appear to have helped the Ss to benefit from the feedback training in Set II, G_2.

During the H trials, both groups showed a significant difference between Sets I and III (Table 17). Estimates of $\underline{w}^2$ revealed that 27 percent of the variance was accounted for by the feedback training on H_{rev} and five percent was accounted for on H_{strt}. Thus, while the H_{rev} group got less out of feedback training on G_2 than did the H_{strt} group, the reverse was true in the case of the respective reinforcement series of H.

The written statements of the two groups are broken down in Table 18. For an explanation of the column headings of Table 18, see the Results section of Experiment G_3.

Table 18

Distribution of Approaches Used by Ss in Experiment H-G as Determined From Written Statements

	Inverted U	Linear	Retrieval	Systematic	Guess	Vague
H_{rev}	3 (9%)	8 (25%)	6 (19%)	2 (6%)	7 (22%)	6 (19%)
H_{strt}	9 (21%)	13 (31%)	1 (2%)	4 (10%)	8 (20%)	6 (15%)

The major differences between the two groups are seen in relation to those who discovered the inverted U and those who used retrieval. It seems that those having had H_{rev} were less likely to discover the twin cutoff points and more likely to be stymied in their endeavours to discover good methods of predictions. This led to more reliance on an inefficient method such as memory.

Table 19 is a summary chart of the experiments in this investigation.

Table 19

Prediction Study Summary Chart

Experiment Title		Preliminary Studies		A
Subsections		1	2	
Purpose		To iron out administrative problems and see if Ss understand task.		To see if Ss improve in ability to predict after feedback training.
Type of Problem		Linear relationship. Predict pass or fail after hearing a score.		Predict pass or fail on basis of pairs of scores.
Number and Type of Subjects		5 friends.	5 introductory psychology students.	6 introductory psychology students.
Variables	Predictor	SCAT "T"*		SCAT "T" H.S.R. "T"
	Predicted	Pass-fail based on 1st semester G.P.A.		Pass-fail based on 1st semester G.P.A.
Item Distribution Across the 3 Sets		45-45-45. All items different.		50-100-50. Sets I and II contain same items. These 50 also intermixed in Set II.
Form of Administration		Oral. Ss respond on cards. Set I was initial test of Ss' ability before training. Set II was feedback reinforcement. Set III was test of training received. Set II reinforcements were on back of Set II response cards.		Two runs: 1st, oral; 2nd, tape with I.B.M. answer sheets. Runs staged 24 hours apart.
Statistical Tests Used		None	None	None
Results and Conclusions		Ss understood and complied with instructions.		Some improvement on Set III over Set I, but not a great deal.

---

* All test items were selected from the records of all freshmen entering the University of Missouri in the fall 1959 semester.

Experiment Title	B	C
Subsections		
Purpose	To see if increasing size of reinforcement set improves Ss' ability to predict.	To determine whether altering the distribution of scores will force Ss to pay more attention to the reinforcement training.
Type of Problem	Predict pass or fail on basis of pair of scores.	Predict pass or fail on basis of pair of scores.
Number and Type of Subjects	10 introductory psychology students.	8 introductory psychology students.
Variable Predictor	SCAT "T" H.S.R. "T"	As Ex. B except all SCAT and H.S.R. scores reduced by 20.
Predicted	Pass-fail based on 1st semester G.P.A.	Pass-fail based on 1st semester G.P.A.
Item Distribution Across the 3 Sets	50-150-50. 50 new items added to Set II of Ex. A.	50-150-50.
Form of Administration	2 runs, 24 hrs. apart. Both on tape. I.B.M. answer sheets.	2 runs, 24 hrs. apart. Both on tape. I.B.M. answer sheets.
Statistical Tests Used	t ω^2	t ω^2
Results and Conclusions	Largest difference in test sets was between Sets I and III of Run I ($t=2.28$, $p<0.025$, $\omega^2=0.17$). Set II enabled the Ss to improve predictive accuracy, though they were still 4 points below level attained using optimum cutoff points.	Largest difference seen was between first and last test sets, but this difference was not significant ($t=1.02$, $\omega^2=0.01$). Altering the scores as done in this experiment did not force the Ss to get more information out of Set II. Apparently the Ss were able to make adequate judgements right from the beginning.

Experiment Title		D
Subsections		
Purpose		To compare a group of naive subjects against a professional one, and to compare both groups with statistical methods of prediction.
Type of Problem		Predict pass or fail on basis of a pair of scores.
Number and Type of Subjects		33 introductory psychology students. 14 experienced counselors and counseling students.
Variables	Predictor	As Experiment C except all scores above 35 increased by their difference from 35 (e.g. 40 became 45).
	Predicted	Pass-fail based on 1st semester G.P.A.
Item Distribution Across the 3 Sets		50-150-50.
Form of Administration		2 runs, 24 hrs. or more apart. I.B.M. answer sheets.
Statistical Tests Used		t ω^2
Results and Conclusions		1. When the results of both groups of Ss were pooled a significant increase was seen over the course of the 2 runs ($t=3.62$, $p<0.005$, $\omega^2=0.11$). Set II contributed significantly to improvement in prediction over the course of the experiment. 2. No significant difference was seen between naive and professional groups on either the first or last given test sets. Nor was there any difference seen between the 2 groups in the amount of improvement over the course of the experiment. Therefore, previous professional counseling experience does not appear to give any advantage in ability to learn in this immediate feedback situation. 3. The Ss did somewhat better than a regression equation based on a fresh sample of 150 items but poorer than the maximum possible when consistent cutoff lines are used (37.23 vs. 35 and 40 respectively).

Experiment Title	E		F
Subsections	1	2	
Purpose	To try out feedback training on a new type of task.	Test limits of this task by giving 2 trials.	To give Ss a more complex task by asking them to make finer discriminations in their predictions.
Type of Problem	Predict whether students who enrolled in engineering stayed in this pattern or transferred to a different one.		The Ss were asked to place the persons who obtained each pair scores into one of five categories.
Number and Type of Subjects	8 intro. psych. students.	7 intro. psych. students.	35 intro. psych. students.
Variables Predictor	MEPT "T", SCAT "T" Verbal SCAT "T" Quantitative, H.S.R. "T".		Same as Experiment D.
Predicted	Transfer or non-transfer from engineering pattern.		A, B, C, D or E based on 1st semester G.P.A.
Item Distribution Across the 3 Sets	30-50-30.		50-150-50 as Experiment D.
Form of Administration	One run. Visual administration I.B.M. answer sheets.	Two runs 24 hrs. apart. Visual administration I.B.M. answer sheets.	One run on tape. I.B.M. answer sheets.
Statistical Tests Used	None	t ω^2 χ^2	t ω^2 χ^2
Results and Conclusions	Subsection 1: Little difference seen between Sets I and III. Ss' ability to predict was not improved by feedback training of Set II. Subsection 2: A significant improvement was seen over the course of the 2 trials ($t=2.29$, $p<0.05$, $\omega^2=0.23$). A test of Ss' probability matching with the actual distribution of N and T items shows that the Ss did not probability match on each administration of Set I but did so on both administrations of Set III ($p>0.90$ and $p>0.99$ on Runs I and II respectively). It was concluded that since there were no discriminable differences between the N and T groups on any of the 4 predictor variables the naive source of improvement was due to better probability matching.		1. Ss improved significantly over the course of the experiment ($t=5.0$, $p<0.001$ $\omega^2=0.26$). Ss did considerably poorer than optimum consistent cutoff points, but somewhat better than regression equations. 2. Subjects improved somewhat but were unable to closely match the actual distribution of A, B, C, D and E items.

Experiment Title	G		
Subsections	1	2	3
Purpose	To see if Ss could discover or learn a non-linear distribution, specifically, an inverted U distribution.		To compare naive Ss and experienced (professional) counselors on the inverted U task.
Type of Problem	Ss were given a distribution where the two extremes tended to have the same outcome while scores in the middle usually differed from this.		
Number and Type of Subjects	9 intro. psych. students.	5 intro. psych. students.	35 intro. psych. students. 28 college counselors.
Variables Predictor	SCAT "T".		
Predicted	Pass or fail. Extremes tended to fail. Middle tend to pass.		
Item Distribution Across the 3 Sets	30-60-30.	30-120-30.	30-180-30.
Form of Administration	1 run on tape. I.B.M. answer sheets.		
Statistical Tests Used	None	None	t ω^2
Results and Conclusions	Subsection 1: Some improvement in mean number of correct predictions, but little evidence that Ss caught on to form of distribution. Only one appeared to discover the inverted U. Subsection 2: Ss actually did slightly worse in Set III than in Set I, although 3 of the 5 appeared to have at least some glimmering of what was required. Subsection 3: Both naive and experienced groups were able to significantly increase the mean number of correct predictions over the course of the experiment.(Naive: $t=6.03$, $p<0.001$, $\omega^2=0.34$, 15 caught on. Professional: $t=2.68$, $p<0.01$, $\omega^2=0.10$, 7 caught on). The naive group improved to a greater degree than did the professional group. However, little difference was seen between the mean number of correct predictions of each group in either Set I or Set III. Both groups were somewhat poorer in the overall mean number correct than that obtainable by the use of optimum cutoff points.		

Experiment Title	H - G
Subsections	H_{rev}-G_2 H_{strt}-G_2
Purpose	To discover whether an unusual distribution given before G_2 would better prepare Ss for the inverted U distribution than would a usual (linear) distribution.
Type of Problem	Control group received a regular linear distribution (H_{strt}) followed by G_2 from the previous experiment. Experimental group (H_{rev}) received an inverted linear distribution (high scores failed, low passed) followed by G_2.
Number and Type of Subjects	41 introductory psychology students in control and 38 in experiment group.
Variables Predictor	SCAT "T" in all cases.
Predicted	Pass-fail based on 1st semester G.P.A. In case of H_{rev} usual case was reversed, pass became fail and fail, pass.
Item Distribution Across the 3 Sets	30-60-30 (H) followed by 30-120-30 (G_2).
Form of Administration	1 run on tape. I.B.M. answer sheets.
Statistical Tests Used	t ω^2
Results and Conclusions	1. Only the control group showed improvement in G_2 ($t=2.62$, $p < .01$, $\omega^2=0.07$). Experimental group t was 1.60 while ω^2 was 0.02. No significant difference was seen between experimental and control groups on either Sets I or III of the G_2 task. The outcome was not as predicted, the experimental group did not show a greater improvement than did the control group. Written comments showed that about three times more H_{strt} Ss than H_{rev} discovered the inverted U distribution. 2. Regarding the respective H tasks, both groups showed a significant improvement between Sets I and III. As expected, the experimental group improved to a greater degree ($t=5.02$, $p<0.001$, $\omega^2=0.27$) than did the control group ($t=2.26$, $p< 0.025$, $\omega^2=0.05$). The H_{rev} task did prove to be a more difficult task than did H_{strt}, but it did not lead to a greater improvement in G_2.

Conclusions and Implications

In general our subjects, undergraduate students and graduate students in counseling psychology, were very good at this prediction task under conditions of immediate feedback. When immediate knowledge of results was not available there were no changes in accuracy. The subjects apparently bring some sets to the task, e.g. to consider high numbers as favorable, which help the prediction in some situations, but hinder change to a different situation. They are able to detect curvilinear as well as rectilinear relationships in the data, but they are slower to detect these. We began with two category predictions, but the subjects did equally well when the criterion was distributed across five categories. The professional experience of graduate students in their internship was apparently of little help in this task. In addition to being no more accurate the graduate students complained more about not knowing the meaning of the scores in advance.

Very few of the subjects reacted as if they were being conditioned or were trying to memorize the situation by rote. Most could verbalize the strategy they were attempting clearly enough so that their rules could be applied by others at their instruction or programmed for computer processing. The two most common strategies were the use of multiple cutoff points and a process resembling the use of a regression equation.

Under certain conditions our undergraduate subjects could be said to perform better at this prediction task than a regression equation. A multiple regression equation assumes that a new sample will be distributed similarly to the one on which the equation is computed or "trained." It further assumes that the predictor variables are compensating, that is that a high score on one and a low score on a second predictor are fully equivalent to two medium range scores whose weighted sums are the same. In previous contest type research pitting regression methods against clinical prediction the regression has been computed under ideal conditions, or at least more ideal than the clinicians had, over larger numbers of cases than the clinicians were allowed for practice. In this investigation a regression equation was computed over a sample similar to that used by the subjects making clinical predictions drawn from the same population and of the same number of cases. The subjects and the regression equation were then tested on the same test sample. Under these conditions the clinical predictions are about as accurate as regression. This appears to be partly due to the fact that the situation was not ideal for regression prediction since the criterion was not continuous. Partly it is due to the better prediction obtained by a multiple cutoff procedure which does not allow high scores on one variable to compensate for low scores on another. However, the best cutoff strategies could be programmed, then the consistency of the computer or rigid application of cutoff rules would result in superior performance. The most valuable aspect of this procedure is that it might be used in other situations to discover better strategies for prediction in other situations.

Most of this investigation was spent studying relationships between two predictor variables and complex criteria. One study was done in which there were four predictors and no regression relationship. The result was probability matching by the subjects on the frequency in each criterion category. The frequencies were much more accurate than in any of the other experiments where attention was apprently directed toward finding some rule of prediction. This was not pursued further because one of our major assumptions was that to make inferences as to method employed by subjects it is necessary to employ data where the relationships can be exhaustively known in advance by the experimenter. We suspect that probability matching is employed in cases where the subject can discover no consistent relationship.

We had hoped to make use of multiscale tests such as the MMPI as predictor variables before the end of the contract, but there was not sufficient time. Research of this type will be quite expensive if proper procedures are used in data collection (Sines, 1964). Another approach which might be fruitful would be the investigation of personality variables between subjects who learn this task easily and those who do not seem to catch on.

Another aspect which deserves attention is the set which each subject brings with him. Oskamp (1961) and Newton (1964) both used a procedure which could be described as inducing a set to approach the task in a particular manner. The effect of various instructional sets could be analyzed.

References

Apostolakos, P. C. Multivariate statistical analysis of cognitive and noncognitive factors and use of the analysis in an actuarial-clinical prediction study. Unpublished Ph.D. Thesis, University of Minnesota, 1957.

Barron, F. Some test correlates of response to psychotherapy. *J. consult. Psychol.*, 1953, *17*, 235-41.

Blenkner, M. Predictive factors in the initial interview in family casework. *Social Service Review*, 1954, *28*, 65-73.

Bloom, R. F. & Brundage, E. G. Prediction of success in elementary schools for enlisted personnel. In Stuit, D. B., *Personnel Research and Test Development in the Bureau of Navel Personnel*. Princeton, N. J.: Princeton University Press, 1947, pp. 233-61.

Bobbitt, J. M. & Newman, S. H. Psychological activities at the United States Coast Guard Academy. *Psychol. Bull.*, 1944, *41*, 368-79.

Borden, H. G. Factors for predicting parole success. *J. of Amer. Inst. of Criminal Law and Criminology*, 1928, *19*, 328-36.

Bradley, A. D. & Stein, Carroll. Counselor experience and accuracy of prediction of client behavior: Rehabilitation. *Pers. and Guid. J.*, 1965, *43*, 1015-1017.

Brown, F. G. 1959-60 Official norms and sectioning procedures for the University of Missouri freshmen placement test battery. University of Missouri Testing and Counseling Service Reports. Columbia, Mo.: Vol. 13, No. 5, June 1959 (Mimeo).

Chauncey, H. Personal communication to Meehl. Cited in Meehl, P. E. *Clinical vs. statistical prediction*. Minneapolis: University of Minnesota Press, 1954.

Cline, V. B. Ability to judge personality assessed with a stress interview and sound film technique. *J. abnorm. soc. Psychol.*, 1955, *50*, 183-187.

Cline, V. B. & Richards, J. M. Jr. Components of accuracy of interpersonal perception scores and the clinical and statistical prediction controversy. *Psychol. Rec.*, 1962, *12*, 373-381.

Conrad, H. S. & Satter, G. A. Use of test scores and quality classification ratings in predicting success in electrician's mates school. *OSRD* report No. 5667, Sept. 3, 1945.

Crow, W. J. The effect of training upon accuracy and variability in interpersonal perception. *J. abnorm. soc. Psychol.*, 1957, *55*, 355-359.

DeGroot, A. D. Via clinical to statistical prediction. *Acta. psychol. Amst.*, 1961, *18*, 274-284.

Dunlap, J. W. & Wantman, M. J. *An investigation of the interview as a technique for selecting aircraft pilots.* Washington, D. C.: Civil Aeronautics Administration. Report No. 33, 1944.

Easterbrook, J. A. The effect of emotion on cue utilization and the organization of behavior. *Psychol. Rev.*, 1959, *66*, 183-201.

Edwards W. Probability learning in 1000 trials. *J. exp. Psychol.*, 1961, *62*, 385-94.

Eells, K. How effective is differential prediction in three types of college curricula? *Educ. psychol. Measmt.*, 1961, *21*, 2, 459-471.

Freehill, M. F. Student self-estimates as guidance in selecting courses. *Coll. & Univ.*, 1952, *27*, 233-242.

Glaser, D. A reconsideration of some parole prediction factors. *Amer. sociol. Rev.*, 1954, *19*, 335-341.

Goldberg, L. R. The effectiveness of clinicians' judgments: the diagnosis of organic brain damage from the Bender-Gestalt Test. *J. consult. Psychol.*, 1959, *23*, 25-33.

Gough, H. G. Clinical versus statistical prediction in psychology. In Postman, L. *Psychology in the making*. New York: Alfred A. Knopf, 1963.

Grebstein, L. C. Relative accuracy of actuarial prediction, experienced clinicians and graduate students on a clinical judgment task. *J. consult. Psychol.*, 1963, *27*, 127-132.

Halbower, C. C. A comparison of actuary versus clinical prediction to classes discriminated by the Minnesota Multiphasic Personality Inventory. *Dissertation Abstr.*, 1955, *15*, 115.

Hammond, K. R. Probabilistic functioning and the clinical method. *Psychol. Rev.*, 1955, *62*, 255-262.

Hammond, K. R., Hursch, Carolyn J., and Todd, F. J. Analyzing the components of clinical inference. *Psychol. Rev.*, 1964, *71*, 438-456.

Hayes, W. L. *Statistics for psychologists.* New York: Holt, Rinehart , and Winston, 1963.

Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychol. Bull.*, 1960, *57*, 116-131.

Hoffman, P. J. Assessment of the independent contributions of predictors. *Psychol. Bull.*, 1962, *59*, 77-80.

Holt, R. R. Clinical and statistical prediction: a reformulation and some new data. *J. abnorm. soc. Psychol.*, 1958, *56*, 1-12.

Holtzman, W. H. Can the computer supplant the clinician? *J. clin. Psychol.*, 1960, *16*, 119-122.

Holtzman, W. H. & Sells, S. B. Prediction of flying success by clinical analysis of test protocols. *J. abnorm. soc. Psychol.*, 1954, *49*, 485-490.

Hovey, H. B. & Stauffacher, J. C. Inductive versus objective prediction from a test. *J. clin. Psychol.*, 1953, *9*, 349-51.

Hursch, C. J. Hammond, K. R. & Hursch, J. L. Some methodological considerations in multiple-cue probability studies. *Psychol. Rev.*, 1964, *71*, 42-60.

Hutt, M. L. Actuarial and clinical approaches to psychodiagnosis. *Psychol. Rep.*, 1956, *2*, 413-419.

Imig, C. The effect of training on clinical prediction. Unpublished doctoral dissertation, University of Missouri, 1965.

Jones, R. L. & Siegel, L. The individual high school as a predictor of college academic performance. *Educ. psychol. Measmt.*, 1962, *22*, 785-789.

Kelly, E. L. & Fiske, D. W. *The selection of clinical psychologists: progress report and preliminary findings.* Ann Arbor, Mich.: Edwards Letter Shop, 1948.

Kelly, E. L. & Fiske, D. W. *The prediction of performance in clinical psychology.* Ann Arbor, Mich.: Univ. of Mich. Press, 1951.

Kleinmuntz, B. & Alexander, L. Computer program for the Meehl-Dahlstrom MMPI profile rules. *Educ. psychol. Measmt.*, 1962, *22*, 193-199.

Lepley, W. M. & Hadley, H. D. Military report. Cited in Super, D. E.: *Appraising vocational fitness.* New York: Harper, 1949.

Lindman, H. & Edwards, W. Supplementary report: Unlearning the gambler's fallacy. *J. exp. Psychol.*, 1961, *62*, 630.

Luft, J. Implicit hypotheses and clinical predictions. J. abnorm. soc. Psychol., 1950, *45*, 756-759.

Mahan, T. W., Jr. Diagnostic consistency and prediction: a note on graduate student skills. *Pers. and Guid. J.*, 1963, *52*, 364-367.

Mann, R. D. A critique of P. E. Meehl's clinical vs. statistical prediction. *Behav. Sci.*, 1956, *1*, 224-230.

McHugh, R. B. & Apostolakos, P. C. Methodology for the comparison of clinical with actuarial predictions. Psychol. Bull., 1959, *56*, 301-308.

Meehl, P. E. *Clinical vs. statistical prediction.* Minneapolis: University of Minnesota Press, 1954.

Meehl, P. E. Wanted - one good cookbook. *Amer. Psychol.*, 1956, *11*, 263-272.

Meehl, P. E. When shall we use our heads instead of the formula? *J. counsel. Psychol.*, 1957, *4*, 268-273.

Meehl, P. E. Some ruminations on the validity of clinical procedures. *Canad. J. Psychol.*, 1959, *13*, 102-128.

Melton, R. S. A comparison of clinical and actuarial methods of prediction with an assessment of the relative accuracy of different clinicians. Unpub. Ph.D. Thesis, University of Minnesota, 1952.

Michael, W. B. & Jones, R. A. Stability of predictive validities of high school grades and of scores on the scholastic aptitude test of the college entrance examination board for liberal arts students. *Educ. psychol. Measmt.*, 1963, *23*, 375-378.

Michael, W. B., Jones, R. A., Cox, Anna, Gershon, A., Hoover, M., Katz, K., and Smith, D. High school record and college board scores as predictors of success in a liberal arts program during the freshman year of college. *Educ. psychol. Measmt.*, 1962, *22*, 399-404.

Muldoon, J. F. & Ray, O. S. A comparison of pattern similarity as measured by six statistical techniques and eleven clinicians. *Educ. psychol.* Measmt., 1958, *18*, 775-781.

Newton, J. Judgment and feedback in a quasi-clinical situation. *J. pers. soc. Psychol.*, 1965, in press.

O'Connell, R. H. Trials with tedium and titillation. Psychol. Bull., 1965, 63, 170-179.

Oskamp, S. W. The relationship of clinical experience and training methods to several criteria of clinical prediction. Unpub. Ph.D. Thesis, Stanford University, 1961.

Posner, M. I. Information reduction in the analysis of sequential tasks. Psychol. Rev., 1964, 71, 491-504.

Prediger, D. J., Krauskopf, C. J., and Callis, R. Predicting academic success at the University of Missouri (for use by faculty advisers). Columbia, Mo.: University of Missouri Testing and Counseling Service Report, Vol. 17, No. 3, Jan. 1963.

Plaut, Erika & Crannell, C. W. The ability of clinical psychologists to discriminate between drawings by deteriorated schizophrenics and drawings by normal subjects. Psychol. Rep., 1955, 1, 153-158.

Sanford, N., McArthur, C. C., Zubin, J., Humphreys, L. G., and Meehl, P. E. Clinical vs. actuarial prediction. Proc. 1955 invit. conf. test. Probl. Educ. Test. Serv. 1956, 93-141.

Sarbin, T. R. The logic prediction in psychology. Psychol. Rev., 1944, 51, 210-28.

Sarbin, T. R. A contribution to the study of actuarial and individual methods of prediction. Amer. J. Sociol., 1942, 48, 593-609.

Schiedt, R. Ein beitrag zum problem der ruckfallsprognose. Ph.D. thesis. Munich: Munchner-Zeitungs-Verlag, 1936. Cited in Meehl, P. E. Clinical vs. statistical prediction. Minneapolis: University of Minnesota Press, 1954.

Seashore, H. G. Women are more predictable than men. J. counsel. Psychol., 1962, 9, 261-270.

Seymour, W. R., Simons, E. N., and Krauskopf, C. J. 1964-65 Official norms, sectioning procedures and trend data for the freshmen placement test battery. Columbia, Mo.: University of Missouri Testing and Counseling Service Report: Vol. 18, No. 4, May, 1964. (Mimeo.)

Sines, J. O. Actuarial methods as appropriate strategy for the validation of diagnostic tests. Psychol. Rev. 1964, 71, 517-523.

Skinner, B. F. Science and human behavior. New York: McMillan Co., 1953.

Stephenson, W. Perspectives in psychology; XX: a note on the methodology of clinical explanation. Psychol. Rec., 1962, 12, 101-103.

Stevenson, H. W. & Zigler, E. F. Probability learning in children. J. exp. Psychol., 1958, 56, 185-192.

Stone, L. A. Relaibility of college freshmen judgements of their own scholastic promise. Psychol. Rep., 1962, 11, 658.

Sydiaha, D. On the equivalence of clinical and statistical methods. J. appl. Psychol., 1959, 43, 395-401.

Taft, R. The ability to judge people. Psychol. Bull., 1955, 52, 1-23.

Thomas, D. S. & Mayo, G. D. A procedure for applying results to the predictions of vocational counselors. Educ. psychol. Measmt., 1957, 17, 416-422.

Truesdell, A. B. & Bath, J. A. Clinical and actuarial predictions of academic survival and attrition. J. counsel. Psychol., 1957, 4, 50-53.

Ullmann, L. P. & Berkman, Virginia C. Judgements of outcome of home cure placement from psychological material. J. clin. Psychol., 1959, 15, 28-31.

Vick, Mary C. & Hornaday, J. A. Predicting grade point average at a small southern college. Ed. Psychol. Measmt., 1962, 22, 759-799.

Ward, J., Jr. Comments on "The paramorphic representation of clinical judgement." Psychol. Bull., 1962, 59, 74-76.

Weinberg, G. H. Clinical versus statistical prediction with a method of evaluating a clinical tool. Dissert. Abstr., 1957, 17, 1602-1603.

Weir, M. W. Developmental changes in problem-solving strategies. Psychol. Rev., 1964, 71, 473-490.

Wirt, R. D. Actuarial prediction. J. consult. Psychol., 1956, 20, 123-124.

Yound, F. C. College freshmen judge their own scholastic promise. Personnel Guid. J., 1954, 32, 399-403.

Appendix

Samples of Pre- & Post-Test Scores and Principles Generated for the Undergraduate Subjects on the Experimental Task

SUBJECT NUMBER UNDERGRADUATE	SESSIONS FIRST PRE-	SESSIONS FIRST POST-	SESSIONS SECOND PRE-	SESSIONS SECOND POST-	PRINCIPLE	PRINCIPLE SCORE
10	35	40	38	37	Both scores above 34 = yes. Both scores below 34 = no. One score above 34 and the other below got a yes or no, depending on their deviations for 34. Example: 25 - 36 = yes. 45 - 30 = no.	37
11	36	35	38	34	The first number seems to show the base score and the second number the amount of improvement. Forty seems to be around the average. If the second number showed a large amount of improvement up to forty, success is indicated. If both numbers are around 40, success is also indicated. If the second number decreased below 40, failure would be indicated.	34

Appendix

CONTINUED

SUBJECT NUMBER	SESSIONS					
	FIRST		SECOND			PRINCIPLE
UNDERGRADUATE	PRE-	POST-	PRE-	POST-	PRINCIPLE	SCORE
20	41	40	36	39	I set an arbitrary average of 34 for the two scores. Any that averaged below this I said no. However, if it was close 30-33, I would say yes when one score was quite a bit larger than the other. For example: 37- 29.	40
21	35	37	35	35	Total two scores = middle 70's. If total a border case (low 70's) placed more emphasis on 1st. number. 39 32 - + 32 39 - 0	39
22	34	37	38	37	Cutting score (weighted composite) of 35 was applied more weight thrown to second number of set. Also some intentional guessing.	39
23	38	39	35	36	I tried to find a minimum number that was a passing score and worked from that.	--